

SCCB Lectures

Johan Suykens

K.U. Leuven, ESAT-SCD/SISTA
Kasteelpark Arenberg 10
B-3001 Leuven (Heverlee), Belgium

Tel: 32/16/32 18 02 - Fax: 32/16/32 19 70
Email: johan.suykens@esat.kuleuven.be
<http://www.esat.kuleuven.be/scd/>

SCCB 2006, Modena Italy, Sept. 2006

SCCB 2006 ◊ Johan Suykens

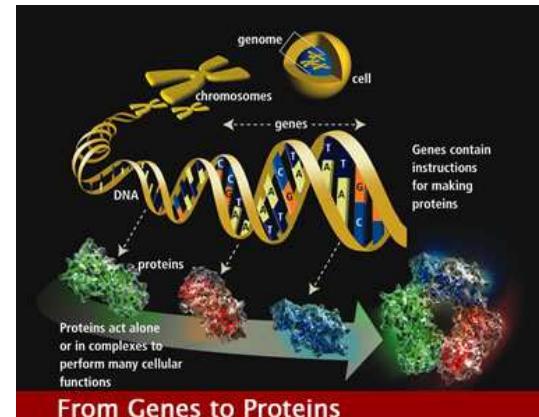


SCCB 2006 ◊ Johan Suykens

Main objectives of the lectures

- Lectures organization:
 1. Support vector machines and kernel based learning (2 x 90 min.)
 2. Case studies (45 min.)
 3. Topics in complex networks, synchronization and cooperative behaviour (90 min.)
- Emphasis on mathematical engineering in multi-disciplinary problems
- Essential concepts
- Providing systematical approaches
- Bridging the gap between theory and practice (coping with fragmentation in science and different fields)
- Understanding different facets of problems

Growth of the “omics”



(From: Human Genome Program, Genomics and its impact on medicine and society, U.S. Department of Energy, 2001)

Genomics (DNAs), **Transcriptomics** (RNAs and gene expression),
Proteomics (protein expression and interactions), **Metabolomics** (metabolic networks), ...

Systems biology



- Molecular biology: decompose system into parts (reductionist approach)
 - Systems biology: integration of parts into a whole (holistic approach)

High-throughput techniques to study genomes and proteomes:

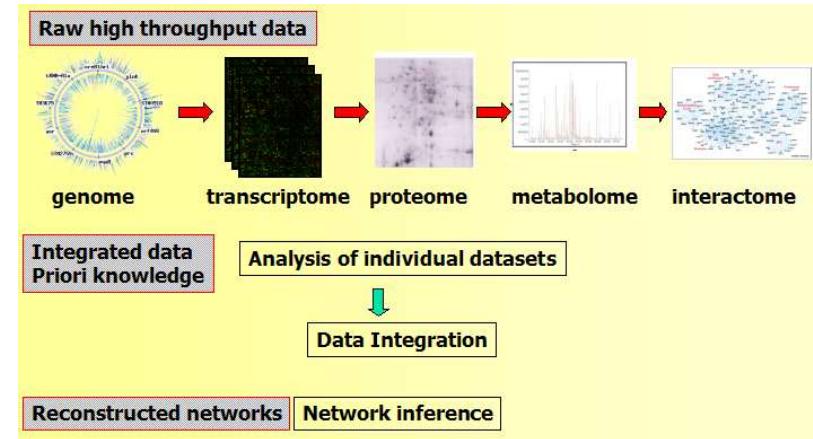
Microarrays to measure changes in mRNAs

Mass spectrometry to identify proteins, quantify protein levels

SCCB 2006 ◊ Johan Suykens

11

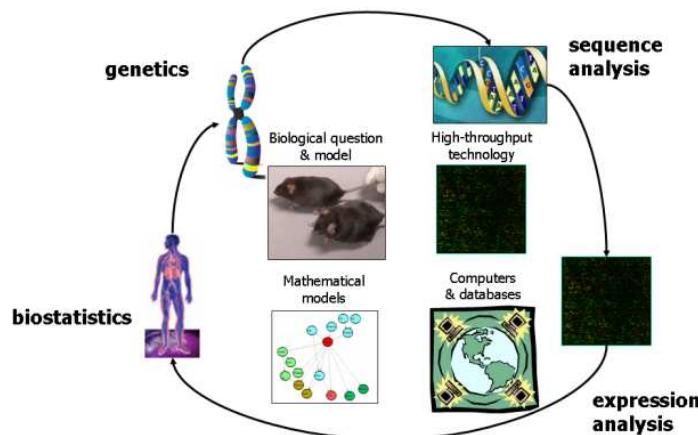
Integrative biology



SCCB 2006 ◊ Johan Suykens

4

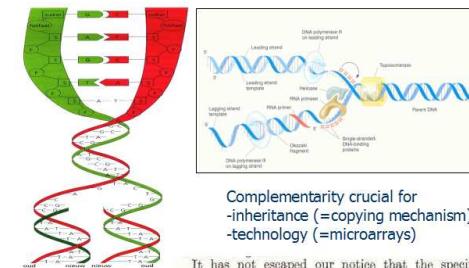
Systems biology (bioinformatics)



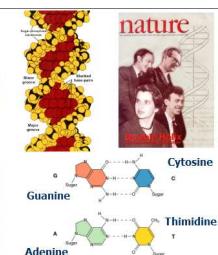
SCCB 2006 ◊ Johan Suykens

(四)

DNA double helix



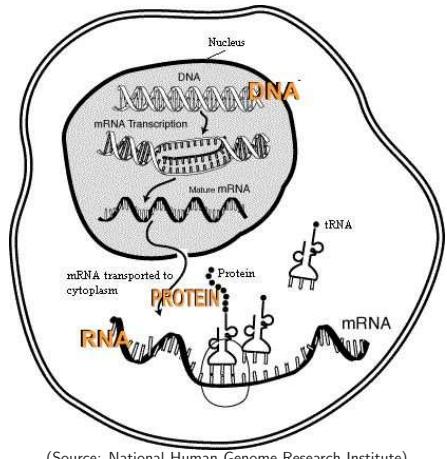
It has not escaped our notice that the specific pairing we have postulated immediately suggests a possibly copying mechanism for the genetic material.



SCCB 2006 ◊ Johan Suykens

6

Central dogma of molecular biology



DNA → RNA → PROTEIN

SCCB 2006 ◊ Johan Suykens

7

Some genomes numbers

Group	Species	Genes	Genome (Mbase)
Phages	Bacteriophage MS2	4	0.003560
Viruses	HIV Type 2	9	0.009671
Bacteria	Haemophilus influenzae (1995)	1760	1.83
Archaea	Methanococcus jannaschii	1735	1.74
Fungi	Saccharomyces cerevisiae (yeast) (1996)	5800	12.1
Protocista	Oxytricha similis	12000	600
Arthropoda	Drosophila melanogaster (fruit fly) (2000)	12000	165
Nematoda	Caenorhabditis elegans (Round worm)(1998)	14000	100
Mollusca	Loligo Pealii	35000	2700
Plantae	Arabidopsis thaliana (Mustard cress)(2000)	25000	70-145
Chordata	Homo Sapiens	30000	3000

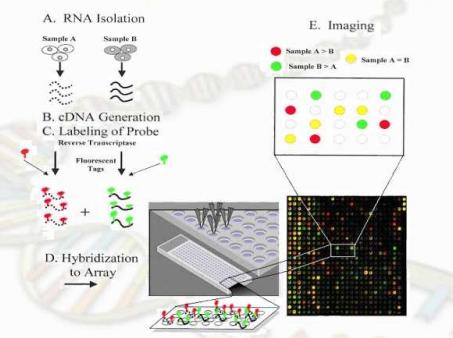
Estimated 265-350 genes are required for 'life'.

SCCB 2006 ◊ Johan Suykens

8

cDNA-microarrays

cDNA microarray array manufacturing

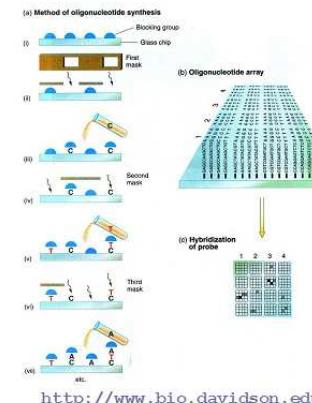


cDNA-microarrays (spotted arrays) (relative measurements, differential hybridization): glass slides on which cDNA is deposited.

SCCB 2006 ◊ Johan Suykens

9

Oligonucleotide microarrays



- clones are covalently bonded on chip
- DNA frames vertically on chip each representative of one gene
- mRNA of test (red tag) and control (green tag) samples
- samples are hybridized to array
- hybridized array scanned by red/green fluorescence

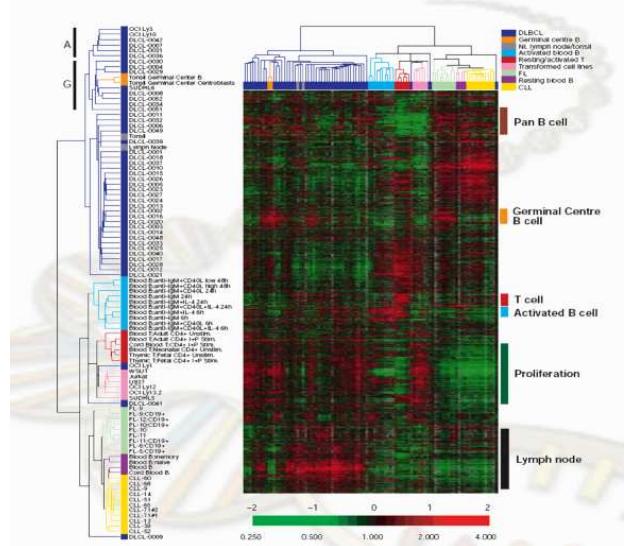
<http://www.bio.davidson.edu/courses/genomics/chip/chip.html>

Oligonucleotide microarrays (DNA chips, Affymetrix) (absolute measurements): produced by the synthesis of oligonucleotides on silicon chips.

SCCB 2006 ◊ Johan Suykens

10

Gene expression data matrix

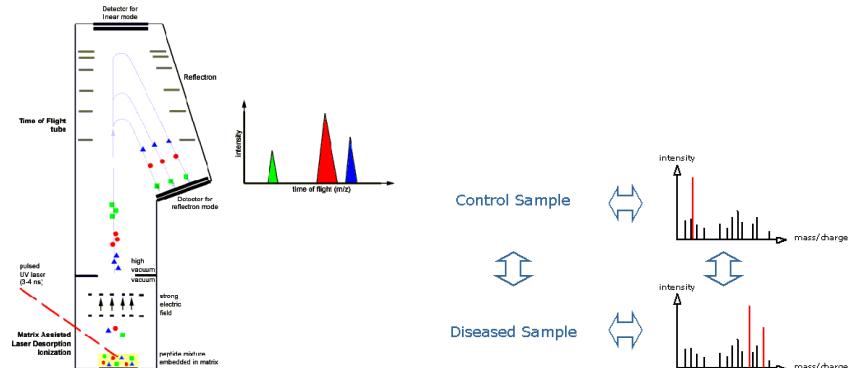


SCCB 2006 ◊ Johan Suykens

11

Proteomics (1)

MALDI-TOF mass spectrometer



Mass Spectrometry: measure the molecular masses of molecules or molecule fragments: mass analysis of complex organic mixtures, identification of proteins and peptides.

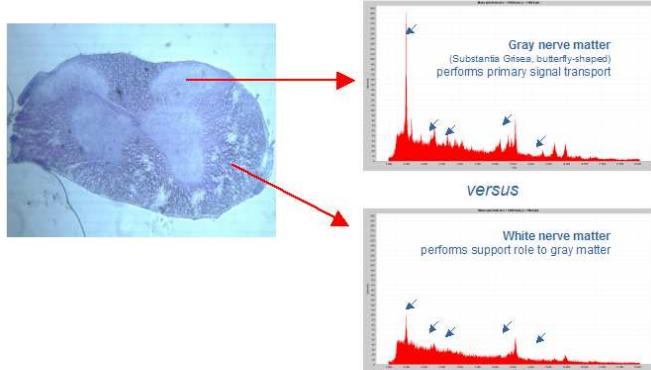
Structural proteomics: X-ray crystallography and NMR spectroscopy (high-throughput determination of protein structures)

SCCB 2006 ◊ Johan Suykens

12

Proteomics (2)

Amyotrophic Lateral Sclerosis (ALS) study



Sample origin: Transversal cross-section of the spinal cord of a standard control rat
(Courtesy of prof. L. Van Den Bosch & M. Dewil, Exp. Neurobiology, K.U.Leuven)

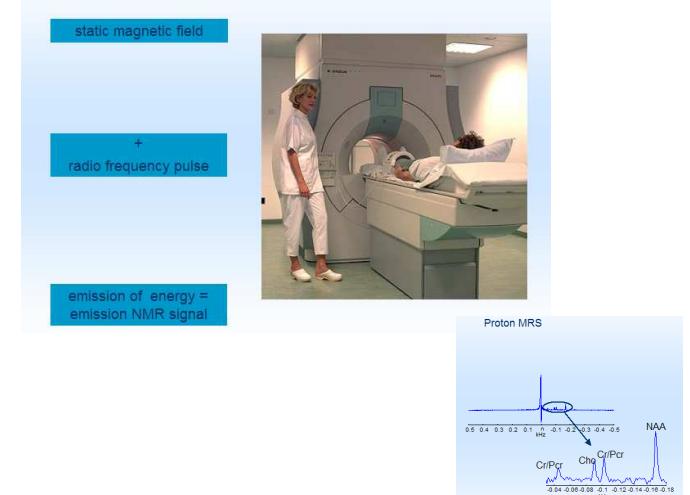
A microscopic image of the nerve tissue section via immunocolouring

(K.U. Leuven Prometa facility www.prometa.kuleuven.be & biomacs.kuleuven.be)

SCCB 2006 ◊ Johan Suykens

13

Nuclear Magnetic Resonance



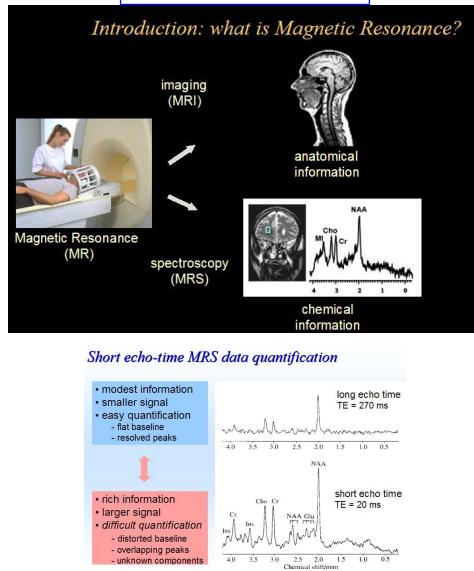
brain tumors, multiple sclerosis, Alzheimer, epilepsy, prostate cancer

SCCB 2006 ◊ Johan Suykens

14

MRI and MRS

Introduction: what is Magnetic Resonance?



SCCB 2006 ◊ Johan Suykens

15

Contents - Part I: Basics

- Motivation
 - Basics of support vector machines
 - Use of the “kernel trick”
 - Kernelbased learning
 - Learning and generalization
 - Least squares support vector machines
 - Primal and dual representations
 - Robustness

Support Vector Machines and Kernel Based Learning

Johan Suykens

K.U. Leuven, ESAT-SCD/SISTA

Kasteelpark Arenberg 10

B-3001 Leuven (Heverlee), Belgium

Tel: 32/16/32 18 02 - Fax: 32/16/32 19 70

Email: johan.suykens@esat.kuleuven.be

<http://www.esat.kuleuven.be/scd/>

SCCB 2006, Modena Italy, Sept. 2006

SCCB 2006 ◊ Johan Suykens

16

Why support vector machines and kernel methods?

- With new technologies (e.g. in microarrays, proteomics) massive data sets become available that are **high dimensional**.
 - Tasks and objectives:** predictive modelling, knowledge discovery and integration, data fusion (classification, feature selection, prior knowledge incorporation, correlation analysis, robustness).
 - Supervised, unsupervised or semi-supervised** learning depending on the given data and problem.
 - Need for modelling techniques that are able to operate on **different data types** (sequences, graphs, numerical, categorical, ...)
 - Linear as well as nonlinear** models
 - Reliable** methods: numerically, computationally, statistically

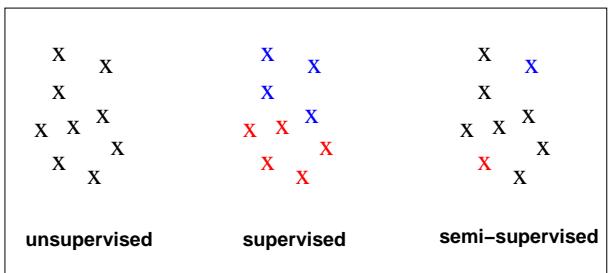
SCCB 2006 ◊ Johan Suykens

17

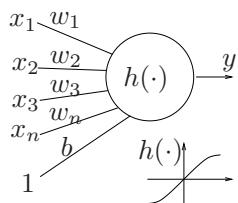
SCCB 2006 ◊ Johan Suykens

18

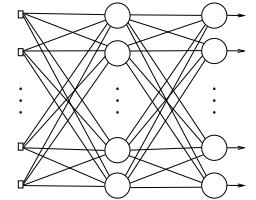
Learning: unsupervised, supervised, semi-supervised



Given data can be labeled, unlabeled or partially labeled
Typically: clustering = unsupervised, classification = supervised



Classical MLPs



Multilayer Perceptron (MLP) properties:

- Universal approximation of continuous nonlinear functions
- Learning from input-output patterns; either off-line or on-line learning
- Parallel network architecture, multiple inputs and outputs

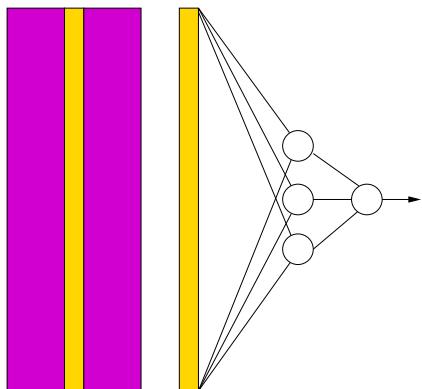
Use in feedforward and recurrent networks

Use in supervised and unsupervised learning applications

Problems: Existence of many local minima!

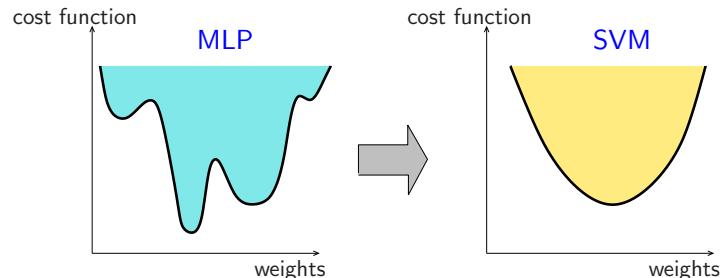
How many neurons needed for a given task?

Classically: need for dimensionality reduction



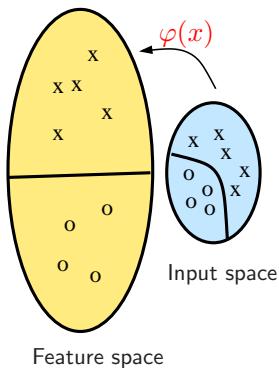
Gene expression matrix ($10.000 \text{ genes} \times 50 \text{ patients}$):
MLP with 3 hidden units would need estimating more than 30.000 weights
→ traditionally: first dimensionality reduction needed (e.g. PCA)
(MLP model is not suitable on very high dimensional input vectors)

Support Vector Machines



- Nonlinear classification and function estimation by **convex optimization** with a unique solution and primal-dual interpretations.
- **Number of neurons** automatically follows from a convex program.
- Learning and generalization in **large dimensional** input spaces (coping with the curse of dimensionality).
- Use of **kernels** (e.g. linear, polynomial, RBF, MLP, splines, ...). Application-specific kernels possible (e.g. textmining, bioinformatics)

SVMs: living in two worlds ...



Primal space: (\rightarrow large data sets)

Parametric: estimate $w \in \mathbb{R}^{n_h}$
 $y(x) = \text{sign}[w^T \varphi(x) + b]$

$$\varphi_1(x)$$

$$y(x)$$

$$w_1$$

$$w_{n_h}$$

$$\varphi_{n_h}(x)$$

Dual space: (\rightarrow high dimensional inputs)

Non-parametric: estimate $\alpha \in \mathbb{R}^N$
 $y(x) = \text{sign}[\sum_{i=1}^{\#sv} \alpha_i y_i K(x, x_i) + b]$

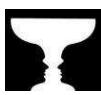
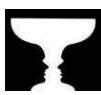
$$K(x, x_1)$$

$$y(x)$$

$$\alpha_1$$

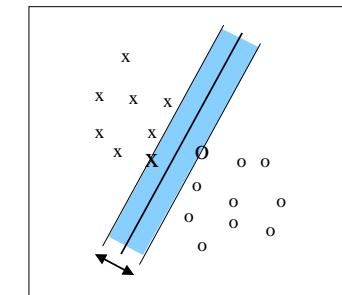
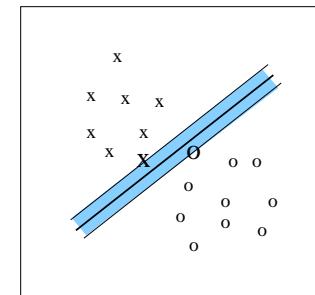
$$\alpha_{\#sv}$$

$$K(x, x_{\#sv})$$



Classifier with maximal margin

- Training set $\{(x_i, y_i)\}_{i=1}^N$: inputs $x_i \in \mathbb{R}^n$; class labels $y_i \in \{-1, +1\}$
- Classifier: $y(x) = \text{sign}[w^T \varphi(x) + b]$
with $\varphi(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{n_h}$ the mapping to a high dimensional feature space (which can be infinite dimensional!)
- Maximize the margin for good generalization ability (margin = $\frac{2}{\|w\|_2}$)
(VC theory: linear SVM classifier dates back from the sixties)



SVM classifier: primal and dual problem

- Primal problem: [Vapnik, 1995]

$$\min_{w, b, \xi} \mathcal{J}(w, \xi) = \frac{1}{2} w^T w + c \sum_{i=1}^N \xi_i \quad \text{s.t.} \quad \begin{cases} y_i [w^T \varphi(x_i) + b] \geq 1 - \xi_i \\ \xi_i \geq 0, \quad i = 1, \dots, N \end{cases}$$

Trade-off between margin maximization and tolerating misclassifications

- Conditions for optimality from Lagrangian.
Express the solution in the Lagrange Multipliers.
- Dual problem: QP problem (convex problem)

$$\max_{\alpha} \mathcal{Q}(\alpha) = -\frac{1}{2} \sum_{i,j=1}^N y_i y_j K(x_i, x_j) \alpha_i \alpha_j + \sum_{j=1}^N \alpha_j \quad \text{s.t.} \quad \begin{cases} \sum_{i=1}^N \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq c, \quad \forall i \end{cases}$$

Obtaining solution via Lagrangian

- Lagrangian:

$$\mathcal{L}(w, b, \xi; \alpha, \nu) = \mathcal{J}(w, \xi) - \sum_{i=1}^N \alpha_i \{y_i [w^T \varphi(x_i) + b] - 1 + \xi_i\} - \sum_{i=1}^N \nu_i \xi_i$$

- Find saddle point: $\max_{\alpha, \nu} \min_{w, b, \xi} \mathcal{L}(w, b, \xi; \alpha, \nu)$, one obtains

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{i=1}^N \alpha_i y_i \varphi(x_i) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow \sum_{i=1}^N \alpha_i y_i = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \rightarrow 0 \leq \alpha_i \leq c, \quad i = 1, \dots, N \end{cases}$$

Finally, write the solution in terms of α (Lagrange multipliers).

SVM classifier model representations

- Classifier: primal representation

$$y(x) = \text{sign}[w^T \varphi(x) + b]$$

Kernel trick (Mercer Theorem): $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$

- Dual representation:

$$y(x) = \text{sign}\left[\sum_i \alpha_i y_i K(x, x_i) + b\right]$$

Some possible kernels $K(\cdot, \cdot)$:

$$K(x, x_i) = x_i^T x \text{ (linear SVM)}$$

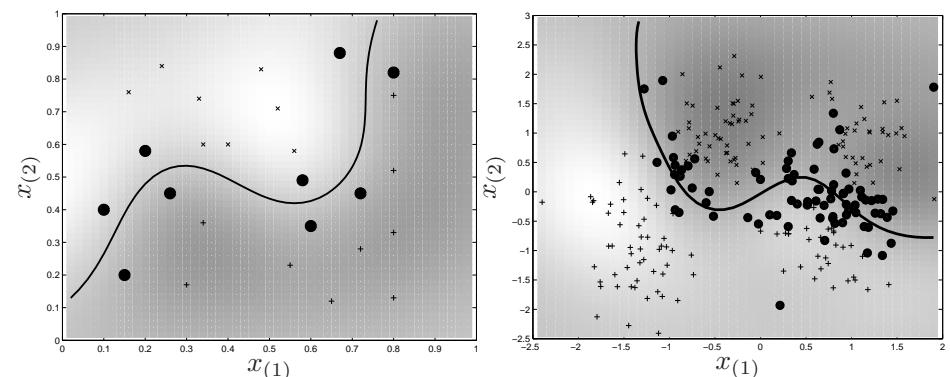
$$K(x, x_i) = (x_i^T x + \tau)^d \text{ (polynomial SVM of degree } d\text{)}$$

$$K(x, x_i) = \exp(-\|x - x_i\|_2^2/\sigma^2) \text{ (RBF kernel)}$$

$$K(x, x_i) = \tanh(\kappa x_i^T x + \theta) \text{ (MLP kernel)}$$

- Sparseness property (many $\alpha_i = 0$)

SVM: support vectors



- Decision boundary can be expressed in terms of a limited number of support vectors (subset of given training data); sparseness property
- Classifier follows from the solution to a convex QP problem.

Reproducing Kernel Hilbert Space (RKHS) view

- Variational problem:** [Wahba, 1990; Poggio & Girosi, 1990; Evgeniou et al., 2000] find function f such that

$$\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda \|f\|_K^2$$

with $L(\cdot, \cdot)$ the loss function. $\|f\|_K$ is norm in RKHS \mathcal{H} defined by K .

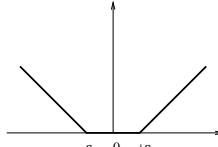
- Representer theorem:** for any convex loss function the solution is of the form

$$f(x) = \sum_{i=1}^N \alpha_i K(x, x_i)$$

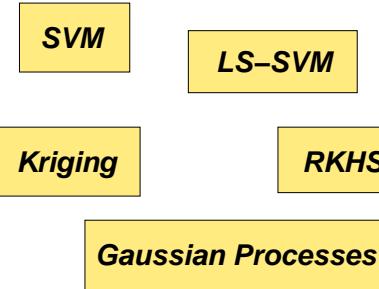
- Some special cases:

$$L(y, f(x)) = (y - f(x))^2 \quad \text{regularization network}$$

$$L(y, f(x)) = |y - f(x)|_\epsilon \quad \text{SVM regression with } \epsilon\text{-insensitive loss function}$$



Different views on kernel machines



Some early history on RKHS:

1910-1920: Moore

1940: Aronszajn

1951: Krige

1970: Parzen

1971: Kimeldorf & Wahba

Obtaining complementary insights from different perspectives:
kernels are used in different settings (try to get the big picture)

Support vector machines (SVM):

Reproducing kernel Hilbert spaces (RKHS):

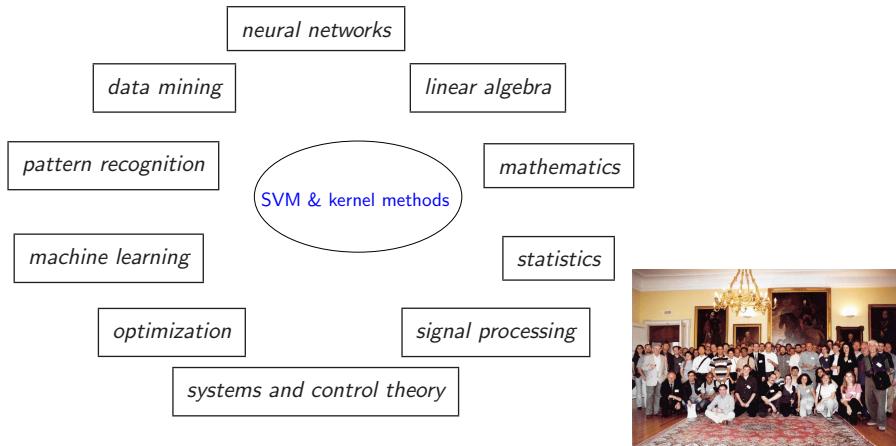
Gaussian processes (GP):

optimization approach (primal/dual)

variational problem, functional analysis

probabilistic/Bayesian approach

Interdisciplinary challenges



NATO Advanced Study Institute on Learning Theory and Practice (Leuven, 2002)

<http://www.esat.kuleuven.ac.be/sista/natoasi/ltp2002.html>

SCCB 2006 ◊ Johan Suykens

31

Wider use of the kernel trick

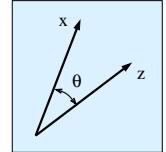
- **Angle between vectors:** (e.g. correlation analysis)

Input space:

$$\cos \theta_{xz} = \frac{x^T z}{\|x\|_2 \|z\|_2}$$

Feature space:

$$\cos \theta_{\varphi(x), \varphi(z)} = \frac{\varphi(x)^T \varphi(z)}{\|\varphi(x)\|_2 \|\varphi(z)\|_2} = \frac{K(x, z)}{\sqrt{K(x, x)} \sqrt{K(z, z)}}$$



- **Distance between vectors:** (e.g. for “kernelized” clustering methods)

Input space:

$$\|x - z\|_2^2 = (x - z)^T (x - z) = x^T x + z^T z - 2x^T z$$

Feature space:

$$\|\varphi(x) - \varphi(z)\|_2^2 = K(x, x) + K(z, z) - 2K(x, z)$$

SCCB 2006 ◊ Johan Suykens

32

Training, validation, test set

- Simplest procedure:

Selection of tuning parameters (regularization constants, kernel tuning parameters) on a validation set, such that one may hope for a good generalization on test data.



- Better: leave-one-out crossvalidation, 10-fold crossvalidation

SCCB 2006 ◊ Johan Suykens

33

Important goal: good generalization

- In fact one would like to minimize the **generalization error**:

$$E[f] = \int_{\mathcal{X} \times \mathcal{Y}} L(y, f(x)) dP(x, y)$$

instead of the **empirical error** (training data) $E_N[f] = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i))$

(with loss function $L(y, f(x))$), i.i.d. samples from a fixed but unknown probability distribution $P(x, y)$ and random variables $x \in \mathcal{X}, y \in \mathcal{Y}$)

- Generalization error bounds, suitable for model selection (if sharp!)
- General message: **avoid overfitting** by taking a good choice of the model complexity (depending on the framework: size of hypothesis space, VC dimension, effective number of parameters, degrees of freedom)
- Occam's razor principle:
“*Entia non sunt multiplicanda praeter necessitatem*”

SCCB 2006 ◊ Johan Suykens

34

Generalization: different mathematical frameworks

- **Vapnik et al.:**

Predictive learning problem (inductive inference)

Estimating values of functions at given points (transductive inference)

Vapnik V. (1998) *Statistical Learning Theory*, John Wiley & Sons, New York.

- **Poggio et al., Smale:**

Estimate true function f with analysis of approximation error and sample error (e.g. in RKHS space, Sobolev space)

Cucker F., Smale S. (2002) "On the mathematical foundations of learning theory", *Bulletin of the AMS*, 39, 1–49.

Poggio T., Rifkin R., Mukherjee S., Niyogi P. (2004) "General conditions for predictivity in learning theory", *Nature*, 428 (6981): 419-422.

Other: Rademacher complexity, stability of learning machines, ...

LS-SVM classifier

- Preserve support vector machine methodology, but simplify via least squares and equality constraints [Suykens, 1999]

- **Primal problem:**

$$\min_{w,b,e} \mathcal{J}(w, e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^N e_i^2 \quad \text{s.t. } y_i [w^T \varphi(x_i) + b] = 1 - e_i, \quad \forall i$$

- **Dual problem:**

$$\left[\begin{array}{c|c} 0 & y^T \\ \hline y & \Omega + I/\gamma \end{array} \right] \left[\begin{array}{c} b \\ \alpha \end{array} \right] = \left[\begin{array}{c} 0 \\ 1_N \end{array} \right]$$

where $\Omega_{ij} = y_i y_j \varphi(x_i)^T \varphi(x_j) = y_i y_j K(x_i, x_j)$ for $i, j = 1, \dots, N$
and $y = [y_1; \dots; y_N]$.

Least Squares Support Vector Machines: "core problems"

- Regression

$$\min_{w,b,e} w^T w + \gamma \sum_i e_i^2 \quad \text{s.t. } y_i = w^T \varphi(x_i) + b + e_i, \quad \forall i$$

- Classification

$$\min_{w,b,e} w^T w + \gamma \sum_i e_i^2 \quad \text{s.t. } y_i (w^T \varphi(x_i) + b) = 1 - e_i, \quad \forall i$$

- Principal component analysis

$$\min_{w,b,e} w^T w - \gamma \sum_i e_i^2 \quad \text{s.t. } e_i = w^T \varphi(x_i) + b, \quad \forall i$$

- Canonical correlation analysis/partial least squares

$$\min_{w,v,b,d,e,r} w^T w + v^T v + \nu_1 \sum_i e_i^2 + \nu_2 \sum_i r_i^2 - \gamma \sum_i e_i r_i \quad \text{s.t. } \begin{cases} e_i = w^T \varphi_1(x_i) + b \\ r_i = v^T \varphi_2(y_i) + d \end{cases}$$

- partially linear models, spectral clustering, subspace algorithms, ...

Obtaining solution from Lagrangian

- **Lagrangian:**

$$\mathcal{L}(w, b, e; \alpha) = \mathcal{J}(w, e) - \sum_{i=1}^N \alpha_i \{y_i [w^T \varphi(x_i) + b] - 1 + e_i\}$$

with Lagrange multipliers α_i (support values).

- **Conditions for optimality:**

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 & \rightarrow w = \sum_{i=1}^N \alpha_i y_i \varphi(x_i) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 & \rightarrow \sum_{i=1}^N \alpha_i y_i = 0 \\ \frac{\partial \mathcal{L}}{\partial e_i} = 0 & \rightarrow \alpha_i = \gamma e_i, \quad i = 1, \dots, N \\ \frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 & \rightarrow y_i [w^T \varphi(x_i) + b] - 1 + e_i = 0, \quad i = 1, \dots, N \end{cases}$$

Eliminate w, e and write solution in α, b .

LS-SVM classifiers benchmarking

- LS-SVM classifiers perform very well on 20 UCI benchmark data sets (10 binary, 10 multiclass) in comparison with many other methods.

but: be aware of the "No free lunch theorem"

	bal	cmc	ims	iri	led	thy	usp	veh	wav	win
NCV	416	982	1540	100	2000	4800	6000	564	2400	118
N _{test}	209	491	770	50	1000	2400	3298	282	1200	60
N	625	1473	2310	150	3000	7200	9298	846	3600	178
nnum	4	2	18	4	0	6	256	18	19	13
ncat	0	7	0	0	7	15	0	0	0	0
n	4	9	18	4	7	21	256	18	19	13
M	3	3	7	3	10	3	10	4	3	3
n _y ,MOC	2	2	3	2	4	2	4	2	2	2
n _{y,1vs1}	3	3	21	3	45	3	45	6	2	3

[Van Gestel et al., Machine Learning 2004]

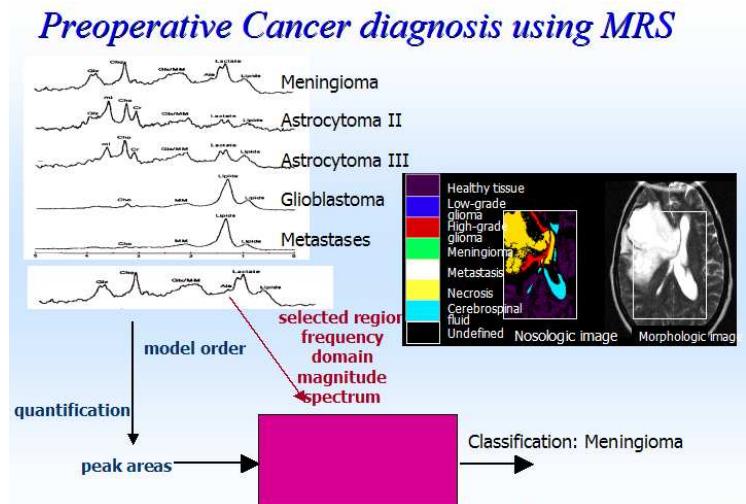
- Winning results in competition WCCI 2006 by [Cawley, 2006]

Benchmarking SVM & LS-SVM classifiers

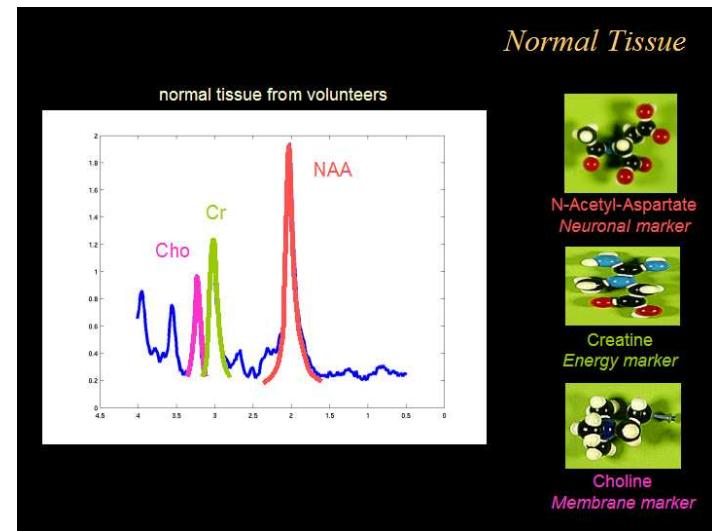
	acr	bld	gcr	hea	ion	pid	snr	ttt	wbc	adu	AA	AR	PST
N _{test}	230	115	334	90	117	256	70	320	228	12222			
n	14	6	20	13	33	8	60	9	9	14			
RBF LS-SVM	87.0 (2.1)	70.2 (4.1)	76.3 (1.4)	84.7 (4.8)	96.0 (2.1)	76.8 (1.7)	73.1(4.2)	99.0 (0.3)	96.4 (1.0)	84.7 (0.3)	84.4	3.5	0.727
RBF LS-SVM _F	86.4 (1.9)	65.1 (2.9)	70.8 (2.4)	83.2 (5.0)	93.4 (2.7)	72.9 (2.0)	73.6 (4.6)	97.9 (0.7)	96.8 (0.7)	77.6 (1.3)	81.8	8.8	0.109
Lin LS-SVM	86.8 (2.2)	65.6 (3.2)	75.4 (2.3)	84.9 (4.5)	87.9 (2.0)	76.8 (1.8)	72.6 (3.7)	66.8 (3.9)	98.5 (1.0)	81.8 (0.3)	79.4	7.7	0.109
Lin LS-SVM _F	86.5 (2.1)	61.8 (3.3)	68.6 (2.3)	82.8 (4.4)	85.0 (3.5)	73.1 (1.7)	73.3 (3.4)	57.6 (1.9)	96.9 (0.7)	71.3 (0.3)	75.7	12.1	0.109
Pol LS-SVM	86.5 (2.2)	70.4 (3.7)	76.3 (1.4)	83.7 (3.9)	91.0 (2.5)	77.0 (1.8)	76.9 (4.7)	99.5 (0.5)	96.4 (0.9)	84.6 (0.3)	84.2	4.1	0.727
Pol LS-SVM _F	86.6 (2.2)	65.3 (2.9)	70.3 (2.3)	82.4 (4.6)	91.7 (2.6)	73.0 (1.8)	77.3 (2.6)	79.1 (0.8)	96.9 (0.7)	77.9 (0.2)	82.0	8.2	0.344
RBF SVM	86.3 (1.8)	70.4 (3.2)	75.9 (1.4)	84.7 (4.8)	95.4 (1.7)	77.3 (2.2)	75.0 (6.6)	98.6 (0.5)	96.4 (1.0)	84.4 (0.3)	84.4	4.0	1.000
Lin SVM	86.7 (2.4)	67.7 (2.6)	75.4 (1.7)	83.2 (4.2)	87.1 (3.4)	77.0 (2.4)	74.1 (4.2)	66.2 (3.6)	96.3 (1.0)	83.9 (0.2)	79.8	7.5	0.021
LDA	85.9 (2.2)	65.4 (3.2)	75.9 (2.0)	83.9 (4.3)	87.1 (2.3)	76.7 (2.0)	67.9 (4.9)	68.0 (3.0)	95.6 (1.1)	82.2 (0.3)	78.9	9.6	0.004
QDA	80.1 (1.9)	62.2 (3.6)	72.5 (1.4)	78.4 (4.0)	90.6 (2.2)	74.2 (3.3)	53.6 (7.4)	75.1 (4.0)	94.5 (0.6)	80.7 (0.3)	76.2	12.6	0.002
Logit	86.8 (2.4)	66.3 (3.1)	76.3 (2.1)	82.9 (4.0)	86.2 (3.5)	77.2 (1.8)	68.4 (5.2)	68.3 (2.9)	96.1 (1.0)	83.7 (0.2)	79.2	7.8	0.109
C4.5	85.5 (2.1)	63.1 (3.8)	71.4 (2.0)	78.0 (4.2)	90.6 (2.2)	73.5 (3.0)	72.1 (2.5)	84.2 (1.6)	94.7 (1.0)	85.6 (0.3)	79.9	10.2	0.021
oneR	85.4 (2.1)	56.3 (4.4)	66.0 (3.0)	71.7 (3.6)	83.6 (4.8)	71.3 (2.7)	62.6 (5.5)	70.7 (1.5)	91.8 (1.4)	80.4 (0.3)	74.0	15.5	0.002
IB1	81.1 (1.9)	61.3 (6.2)	69.3 (2.6)	74.3 (4.2)	87.2 (2.8)	69.6 (2.4)	77.7 (4.4)	82.3 (3.3)	95.3 (1.1)	78.9 (0.2)	77.7	12.5	0.021
IB10	86.4 (1.3)	60.5 (4.4)	72.6 (1.7)	80.0 (4.3)	85.9 (2.5)	73.6 (2.4)	69.4 (4.3)	94.8 (2.0)	96.4 (1.2)	82.7 (0.3)	80.2	10.4	0.039
NB _k	81.4 (1.9)	63.7 (4.5)	74.7 (2.1)	83.9 (4.5)	92.1 (2.5)	75.5 (1.7)	71.6 (3.5)	71.7 (3.1)	97.1 (0.9)	84.8 (0.2)	79.7	7.3	0.109
NB _n	76.9 (1.7)	56.0 (6.9)	74.6 (2.8)	83.8 (4.5)	82.8 (3.8)	75.1 (2.1)	66.6 (3.2)	71.7 (3.1)	95.5 (0.5)	82.7 (0.2)	76.6	12.3	0.002
Maj. Rule	56.2 (2.0)	56.5 (3.1)	69.7 (2.3)	56.3 (3.8)	64.4 (2.9)	66.8 (2.1)	54.4 (4.7)	66.2 (3.6)	66.2 (2.4)	75.3 (0.3)	63.2	17.1	0.002

[Van Gestel et al., Machine Learning 2004]

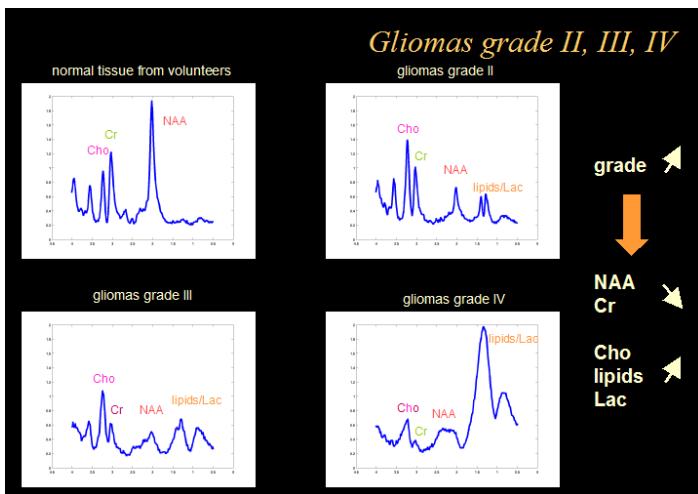
Classification of brain tumors from MRS data (1)



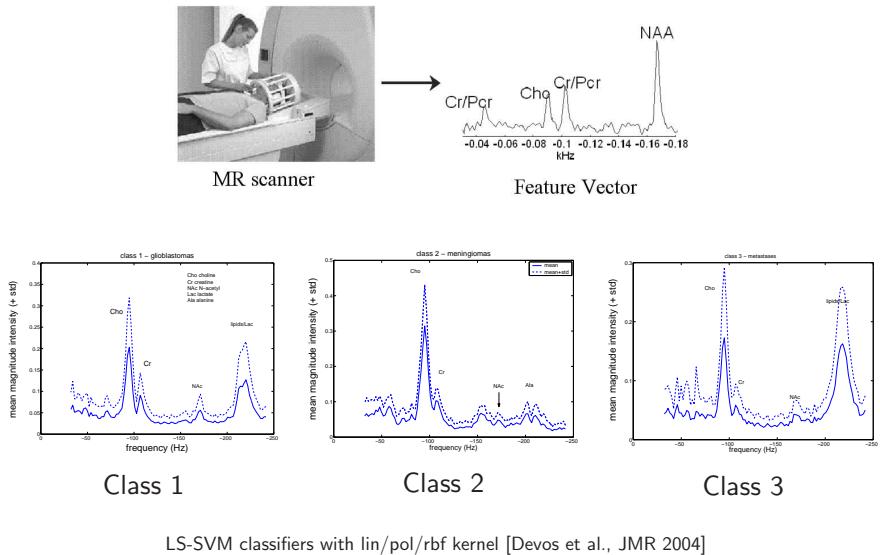
Classification of brain tumors from MRS data (2)



Classification of brain tumors from MRS data (3)



Classification of brain tumors from MRS data (4)



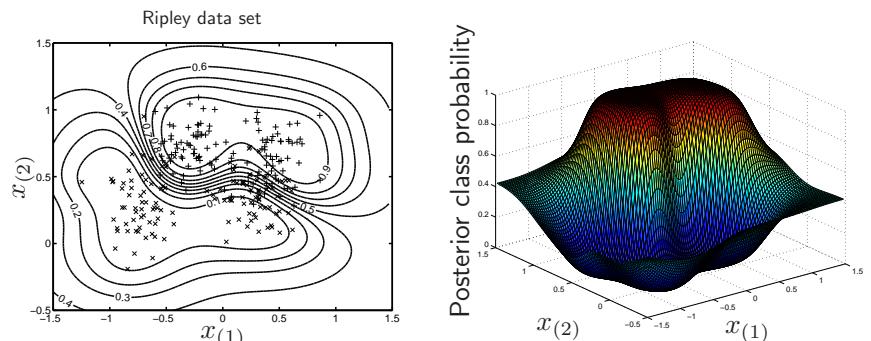
Classification of brain tumors from MRS data (5)

	$e_{train} \pm std(e_{train})$	mean % correct	$e_{test} \pm std(e_{test})$	mean % correct
RBF12	0.0800 ± 0.2727 0.0600 ± 0.2387	99.8621 99.8966	2.8500 ± 1.9968 2.6800 ± 1.6198	90.1724 90.7586
RBF13	1.6700 ± 1.1106 1.7900 ± 1.0473	96.7255 96.4902	8.1200 ± 1.2814 7.7900 ± 1.2815	67.5200 68.8400
RBF23	0 ± 0 0 ± 0	100 100	2.0000 ± 1.1976 2.0200 ± 1.2632	90.4762 90.3810
Lin12, $\gamma=1$	6.2000 ± 1.3333 6.1300 ± 1.4679	89.3100 89.4310	3.8900 ± 1.8472 3.6800 ± 1.7746	86.586 87.3103
Lin13, $\gamma=1$	15.6400 ± 1.7952 15.3700 ± 1.8127	69.333 69.8627	7.6800 ± 0.8863 7.9200 ± 1.0316	69.280 68.3200
Lin23, $\gamma=1$	4.0100 ± 1.3219 4.0000 ± 1.1976	90.452 90.4762	3.4400 ± 1.2253 2.9600 ± 1.3478	83.619 85.9048

Comparison of LS-SVM classification with LOO using RBF and linear kernel, with additional bias term correction ($N_1 = 50, N_2 = 37, N_3 = 26$).

Be careful not to overfit the data (especially with small data sets and nonlinear classifiers)

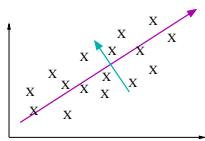
Bayesian inference: classification



- Probabilistic interpretation with moderated output
- Bias term correction for unbalanced and/or small data sets
- Bayesian approaches to kernel methods: Gaussian processes

Classical PCA analysis

- Given zero mean data $\{x_i\}_{i=1}^N$ with $x \in \mathbb{R}^n$
- Find projected variables $w^T x_i$ with maximal variance



$$\begin{aligned}\max_w \text{Var}(w^T x) &= \text{Cov}(w^T x, w^T x) \simeq \frac{1}{N} \sum_{i=1}^N (w^T x_i)^2 \\ &= w^T C w\end{aligned}$$

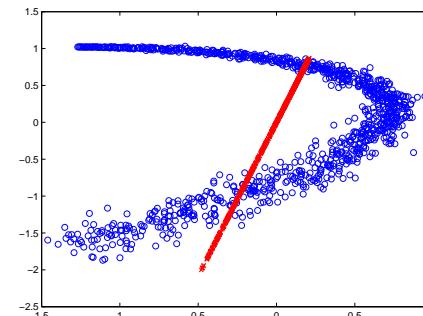
where $C = \frac{1}{N} \sum_{i=1}^N x_i x_i^T$. Consider additional constraint $w^T w = 1$.

- Resulting eigenvalue problem:

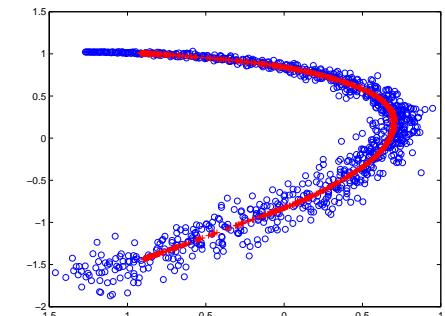
$$Cw = \lambda w$$

with $C = C^T \geq 0$, obtained from the Lagrangian $\mathcal{L}(w; \lambda) = \frac{1}{2} w^T C w - \lambda(w^T w - 1)$ and setting $\partial \mathcal{L}/\partial w = 0$, $\partial \mathcal{L}/\partial \lambda = 0$.

Kernel PCA



linear PCA



kernel PCA (RBF kernel)

[Schölkopf et al., 1998]

Kernel PCA: primal and dual

- Eigenvalue decomposition of the kernel matrix [Schölkopf et al., 1998]
- Primal problem: [Suykens et al., 2003]

$$\min_{w,b,e} \mathcal{J}(w, e) = \gamma \frac{1}{2} w^T w - \frac{1}{2} \sum_{i=1}^N e_i^2 \quad \text{s.t. } e_i = w^T \varphi(x_i) + b, \quad i = 1, \dots, N.$$

- Dual problem = kernel PCA :

$$\Omega_c \alpha = \lambda \alpha \quad \text{with } \lambda = 1/\gamma$$

with $\Omega_{c,ij} = (\varphi(x_i) - \hat{\mu}_\varphi)^T (\varphi(x_j) - \hat{\mu}_\varphi)$ the centered kernel matrix and $\hat{\mu}_\varphi = (1/N) \sum_{i=1}^N \varphi(x_i)$.

- Score variables (allowing also out-of-sample extensions):

$$z(x) = w^T (\varphi(x) - \hat{\mu}_\varphi) = \sum_j \alpha_j (K(x_j, x) - \frac{1}{N} \sum_r K(x_r, x) - \frac{1}{N} \sum_r K(x_r, x_j) + \frac{1}{N^2} \sum_r \sum_s K(x_r, x_s))$$

Obtaining solution from Lagrangian

- Lagrangian (here case $b = 0$)

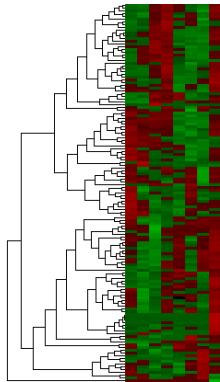
$$\mathcal{L}(w, e; \alpha) = \gamma \frac{1}{2} \sum_{i=1}^N e_i^2 - \frac{1}{2} w^T w - \sum_{i=1}^N \alpha_i (e_i - w^T \varphi(x_i))$$

- Conditions for optimality

$$\left\{ \begin{array}{l} \frac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{i=1}^N \alpha_i \varphi(x_i) \\ \frac{\partial \mathcal{L}}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma e_i, \quad i = 1, \dots, N \\ \frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 \rightarrow e_i - w^T \varphi(x_i) = 0, \quad i = 1, \dots, N \end{array} \right.$$

Eliminate w, e and write solution in α .

Microarray data analysis



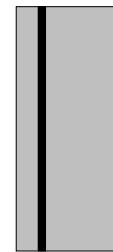
FDA
LS-SVM classifier (linear, RBF)
Kernel PCA + FDA
(unsupervised selection of PCs)
(supervised selection of PCs)

Use regularization for linear classifiers

Systematic benchmarking study in [Pochet et al., Bioinformatics 2004]
Webservice: <http://www.esat.kuleuven.ac.be/MACBETH/>

Primal versus dual problems

Example 1: microarray data (10.000 genes & 50 training data)



Classifier model:
 $\text{sign}(w^T x + b)$ (primal)
 $\text{sign}(\sum_i \alpha_i y_i x_i^T x + b)$ (dual)

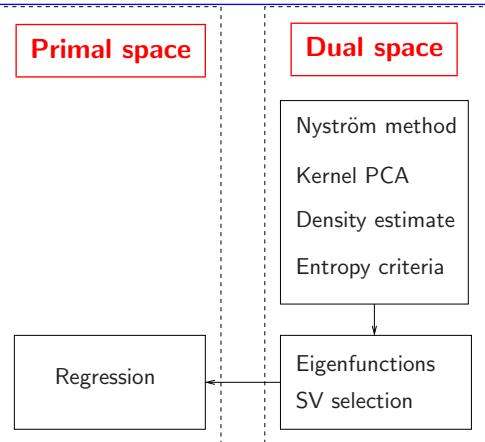
linear FDA primal: $w \in \mathbb{R}^{10.000}$ (only 50 training data!)
 linear FDA dual: $\alpha \in \mathbb{R}^{50}$

Example 2: datamining problem (1.000.000 training data & 20 inputs)



linear FDA primal: $w \in \mathbb{R}^{20}$
 linear FDA dual: $\alpha \in \mathbb{R}^{1.000.000}$ (kernel matrix: $1.000.000 \times 1.000.000$!)

Fixed-size LS-SVM: primal-dual kernel machines



Modelling in view of primal-dual representations

Link Nyström approximation (GP) - kernel PCA - density estimation

[Suykens et al., 2002]: primal space estimation, sparse, large scale

Nyström method (Gaussian processes)

[Williams, 2001 *Nyström method*; Girolami, 2002 *KPCA, density estimation*]

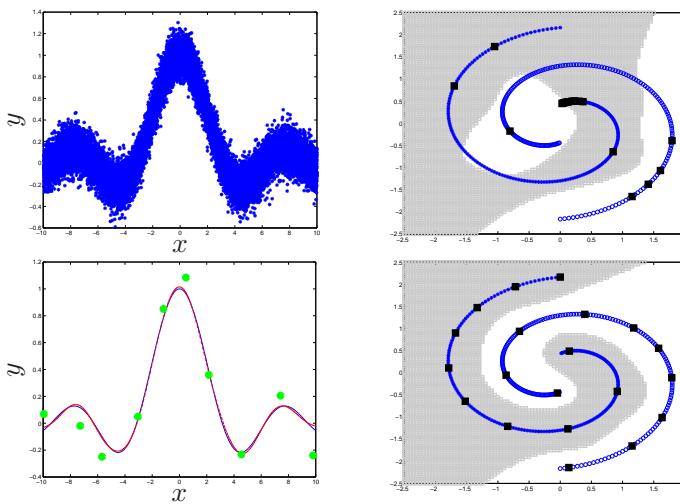
- “big” matrix: $\Omega_{(N,N)} \in \mathbb{R}^{N \times N}$, “small” matrix: $\Omega_{(M,M)} \in \mathbb{R}^{M \times M}$ (based on random subsample, in practice often $M \ll N$)
- Eigenvalue decompositions: $\Omega_{(N,N)} \tilde{U} = \tilde{U} \tilde{\Lambda}$ and $\Omega_{(M,M)} \bar{U} = \bar{U} \bar{\Lambda}$
- Relation to eigenvalues and eigenfunctions of the integral equation

$$\int K(x, x') \phi_i(x) p(x) dx = \lambda_i \phi_i(x')$$

with

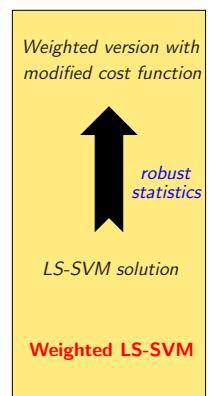
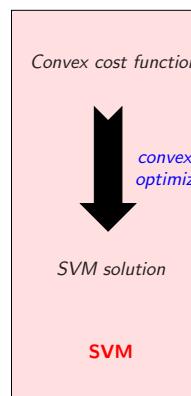
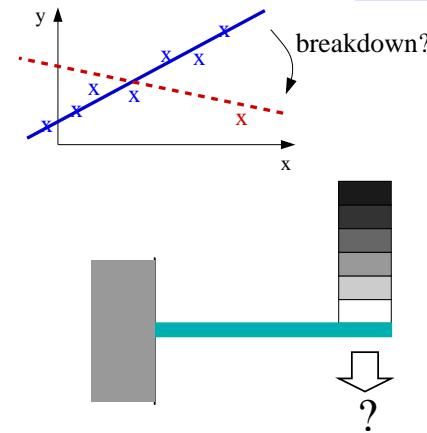
$$\hat{\lambda}_i = \frac{1}{M} \bar{\lambda}_i, \quad \hat{\phi}_i(x_k) = \sqrt{M} \bar{u}_{ki}, \quad \hat{\phi}_i(x') = \frac{\sqrt{M}}{\bar{\lambda}_i} \sum_{k=1}^M \bar{u}_{ki} K(x_k, x')$$

Fixed-size LS-SVM: toy examples



Sparse representations with estimation in primal space

Robustness



Robust statistics: Bounded derivative of loss function, bounded kernel

Linear parametric models: do not start from LS

Kernel based regression: starting from LS is allowed under certain conditions
[Suykens et al., 2002; Debruyne et al., 2006]

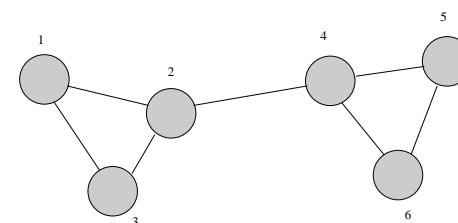
Contents - Part II: Advanced topics

- Spectral graph clustering
- Semi-supervised learning
- Integration of data sources
- Kernels from graphical models
- Kernel canonical correlation analysis
- Sparseness, feature selection, relevance determination
- Prior knowledge incorporation, convex optimization

Graph representation

- Graph representing e.g. a set of proteins:
Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
 $\mathcal{V} = \{x_1, x_2, \dots, x_N\}$: set of vertices (nodes)
 \mathcal{E} : set of edges
 $W = [w_{ij}]$: affinity matrix with similarity values $w_{ij} \geq 0$
 w_{ij} : the more similar x_i and x_j , the larger the value w_{ij}

- Examples: $w_{ij} \in \{0, 1\}$, $w_{ij} = \exp(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2})$



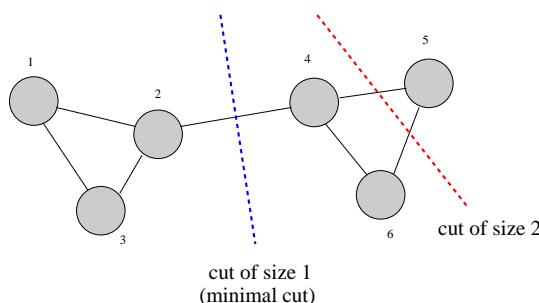
(e.g. $w_{12} = 1, w_{13} = 1, w_{14} = 0$)

Spectral graph clustering (1)

- Discover two clusters \mathcal{A}, \mathcal{B} in the graph \mathcal{G} with **minimal cut**:

$$\min_{q \in \{-1,+1\}^N} \frac{1}{2} \sum_{i,j} w_{ij} (q_i - q_j)^2$$

with **cluster membership indicator** $q_i = 1$ if $i \in \mathcal{A}$, $q_i = -1$ if $i \in \mathcal{B}$.



Spectral graph clustering (2)

- The **min-cut** spectral clustering problem can be written as

$$\min_{q \in \{-1,+1\}^N} q^T (D - W) q$$

with **degree matrix** $D = \text{diag}(d_1, \dots, d_N)$ and degrees $d_i = \sum_j w_{ij}$.

- Relax** the combinatorial problem: $q^T q = 1$ instead of $q \in \{-1,+1\}^N$. This gives the **eigenvalue problem** $L\tilde{q} = \lambda\tilde{q}$ with $L = D - W$ the **Laplacian** of the graph (like kernel PCA on L) [Shi & Malik, 2000; Ng et al. 2002; Chung, 1997].
- Cluster member indicators: $q_i = \text{sign}(\tilde{q}_i - \theta)$ with threshold θ .
- Normalized cut:** $L\tilde{q} = \lambda D\tilde{q}$ (avoids isolated points)
- Note: diffusion kernel $K = \exp(-\beta L)$ [Kondor, 2002]

Underlying primal problems

- Min-cut:** LS-SVM primal problem for kernel PCA

$$\max_{w,b,e} \gamma \frac{1}{2} e^T e - \frac{1}{2} w^T w \text{ such that } e_i = w^T \varphi(x_i) + b, \forall i = 1, \dots, N$$

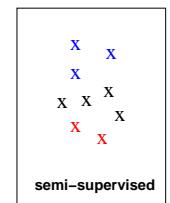
- Normalized:** Weighted LS-SVM primal problem

$$\max_{w,b,e} \gamma \frac{1}{2} e^T V e - \frac{1}{2} w^T w \text{ such that } e_i = w^T \varphi(x_i) + b, \forall i = 1, \dots, N$$

with $V = D^{-1}$ the inverse degree matrix [Alzate & Suykens, 2006]

- Bias term leads to optimal centering.
- Allows for **out-of-sample extensions** on test data and evaluation on validation sets (for tuning parameter selection).

Semi-supervised learning



Semi-supervised learning: part labeled and part unlabeled

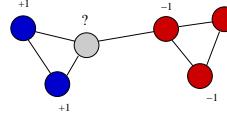
Assumptions for semi-supervised learning to work:

[Chapelle, Schölkopf, Zien, 2006]

- Smoothness assumption:** if two points x_1, x_2 in a high density region are close, then also the corresponding outputs y_1, y_2
- Cluster assumption:** points from the same cluster are likely to be of the same class
- Low density separation:** decision boundary should be in low density region
- Manifold assumption:** data lie on a low-dimensional manifold

Estimation of labels at unlabeled nodes

Functional class prediction on a protein network [Tsuda et al., 2005]



- Total number of nodes: $N = N_l + N_u$:
 N_l labeled nodes with given values $y_1, \dots, y_{N_l} \in \{-1, +1\}$
 N_u unlabeled nodes with values $y_{N_l+1}, \dots, y_{N_l+N_u}$ (assumed 0)
- Goal: find estimated values $\hat{y} = [\hat{y}_1; \dots; \hat{y}_N]$ from

$$\min_{\hat{y}} \sum_{i=1}^{N_l} (\hat{y}_i - y_i)^2 + \mu \sum_{i=N_l+1}^N \hat{y}_i^2 + c \sum_{i,j=1}^N w_{ij} (\hat{y}_i - \hat{y}_j)^2$$

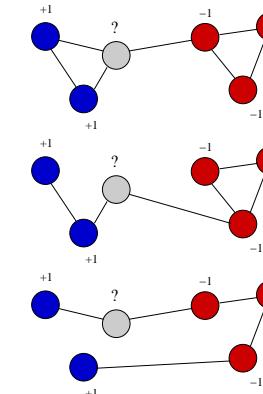
Solution:

$$\hat{y} = (I + cL)^{-1}y$$

with L the Laplacian matrix and $y = [y_1; \dots; y_N]$.

Integration of data sources

Different graphs, each containing a part of information [Tsuda et al., 2005]



Consider linear combination of Laplacians $L = \sum_{j=1}^{N_g} \beta_j L_j$ and solve

$$\min_{\hat{y}, \beta} (\hat{y} - y)^T (\hat{y} - y) + c \hat{y} L \hat{y}$$

Semi-supervised learning in RKHS

- Learning in RKHS [Belkin & Niyogi, 2004]:

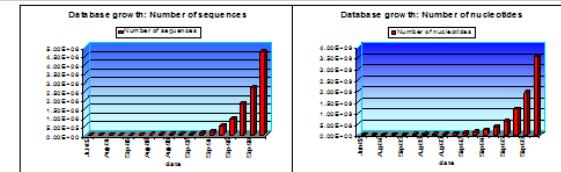
$$\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N V(y_i, f(x_i)) + \lambda \|f\|_K^2 + \eta \mathbf{f}^T L \mathbf{f}$$

with $V(\cdot, \cdot)$ the loss function, L the Laplacian matrix, $\|f\|_K$ is norm in RKHS \mathcal{H} and $\mathbf{f} = [f(x_1); \dots; f(x_{N_l+N_u})]$ (N_l, N_u number of labeled and unlabeled data)

- Laplacian term: discretization of the Laplace-Beltrami operator
- Representer theorem: $f(x) = \sum_{i=1}^{N_l+N_u} \alpha_i K(x, x_i)$
- Least squares solution case: Laplacian acts on kernel matrix

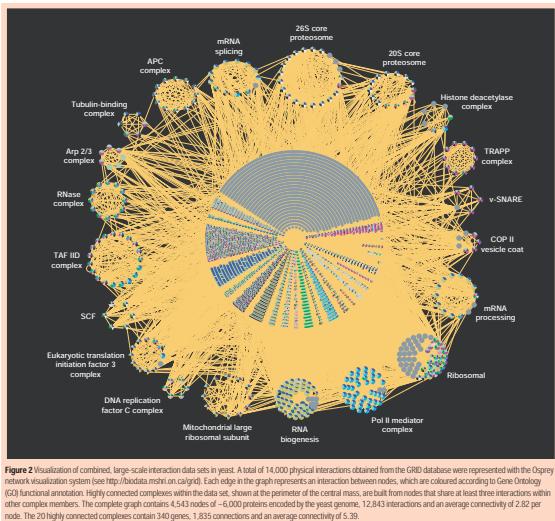
Growth of databases

Database category	Data content	Examples
1. Literature database	Bibliographic citations On-line journals	MEDLINE (1971)
2. Factual database	Nucleic acid sequences Amino acid sequences	GenBank (1982), EMBL (1982), DDBJ (1984) PIR (1968), PRF (1979), SWISS-PROT (1986)
3. Knowledge base	3D molecular structures Motif libraries Molecular classifications Biochemical pathways	PDB (1971), CSD (1965) PROSITE (1988) SCOP (1994) KEGG (1995)



Hence: computational complexity important (e.g. exploit sparse matrices)

Large scale interaction data sets in yeast



[Tyers & Mann, From genomes to proteomics, Nature 2002]

Function class prediction of yeast proteins (1)

- Dataset [Tsuda et al., 2005; Lanckriet et al., 2004]: 3588 proteins
Function of proteins labelled according to MIPS Comprehensive Yeast Genome Database

Focus on 13 highest-level categories of functional hierarchy

Functional classes:

1. metabolism
2. energy
3. cell cycle and DNA processing
4. transcription
5. protein synthesis
6. protein fate
7. cellular transportation
8. cell rescue and defense
9. interaction with cell environment
10. cell fate
11. control of cell organization
12. transport facilitation
13. others

Function class prediction of yeast proteins (2)

- Improved results with **combined Laplacians** [Tsuda et al., 2005]
- Choice of matrices W_i related to the graphs \mathcal{G}_i ($i = 1, 2, \dots, 5$):**

W_1 : Network from Pfam domain structure

W_2 : Co-participation in a protein complex

W_3 : Protein-protein interactions (MIPS physical interactions)

W_4 : Genetic interactions (MIPS genetic interactions)

W_5 : Network created from cell gene expression measurements

(Note: Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families - www.sanger.ac.uk/Software/Pfam/)

Data fusion with kernels

- Consider a **combination of kernel matrices** $K = \sum_{i=1}^m \mu_i K_i$ ($\mu_i \geq 0$) with

Kernel	Data	Similarity measure
K_1	protein sequences	Smith -Waterman
K_2	protein sequences	BLAST
K_3	protein sequences	Pfam HMM
K_4	hydropathy profile	FFT
K_5	protein interactions	linear kernel
K_6	protein interactions	diffusion kernel
K_7	gene expression	RBF kernel
K_8	random numbers	linear kernel

- Improved results by combining kernels [Lanckriet et al., 2004]

Learning the optimal combination

- Learn optimal combination of μ_i together with SVM classifier by solving a single **convex problem** [Lanckriet et al., JMLR 2004]

- QP problem of SVM:

$$\max_{\alpha} 2\alpha^T \mathbf{1} - \alpha^T \text{diag}(y) K \text{diag}(y) \alpha \quad \text{s.t. } 0 \leq \alpha \leq C, \alpha^T y = 0$$

is replaced by

$$\begin{aligned} & \min_{\mu_i} \max_{\alpha} 2\alpha^T \mathbf{1} - \alpha^T \text{diag}(y) \left(\sum_{i=1}^m \mu_i K_i \right) \text{diag}(y) \alpha \\ & \text{s.t. } 0 \leq \alpha \leq C, \quad \alpha^T y = 0, \quad \text{trace} \left(\sum_{i=1}^m \mu_i K_i \right) = c, \quad \sum_{i=1}^m \mu_i K_i \succeq 0. \end{aligned}$$

Can be solved as semidefinite program (**SDP problem**) [Boyd & Vandenberghe, 2004] (LMI constraint for positive definite kernel)

Gene prioritization through genomic data fusion

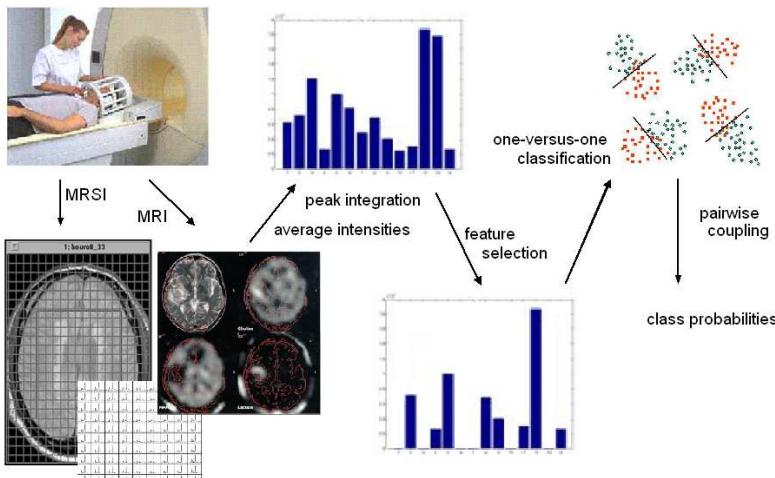
- [Aerts et al., Nature Biotechnology 2006]

Integrating multiple heterogeneous data sources (microarray, BIND, BLAST, cis-regulatory modules, EST, GO, InterPro, KEGG, transcription motifs, literature)

Overall prioritization obtained by data fusion with a global ranking using order statistics.

- The approach successfully identified a novel gene in DiGeorge syndrome (in vivo validation, zebrafish)
- Potential for such methodologies towards kernel methods

Combined MRI and MRS classification system (1)



[Luts et al., 2006]

Combined MRI and MRS classification system (2)

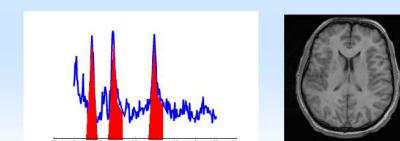
Pre-operative Cancer diagnosis using MRS – Brain tumour classification using MRS & MRI

► provided by UMCN Nijmegen, INTERPRET project (IST-1999-10310)
► 25 brain tumour patients and 4 volunteers, MRI & short echo MRSI

	Label	Pathology	# data	# subject
healthy	1	normal brain tissue	142	4
	2	normal brain patients	76	4
	3	cerebrospinal fluid (CSF)	100	8
	4	gliomas grade II	176	10
	5	gliomas grade III	57	4
	6	glioblastoma	70	7
	7	metastases	48	3
total				669
29				

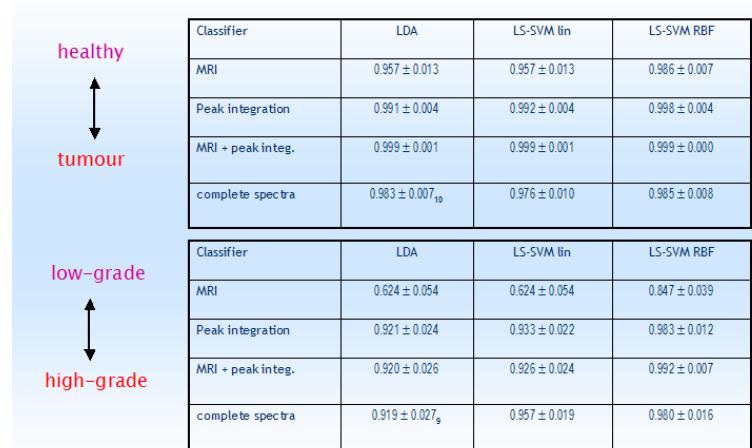
Brain tumour classification using MRS & MRI – Input features

- magnitude spectrum (231 variables)
- peak integration (10): rough estimate of the amplitude
- image information (4): T₁, T₂, PD, GD
- image information + peak integration (14 variables)



Combined MRI and MRS classification system (3)

Brain tumour classification using MRS & MRI--Results

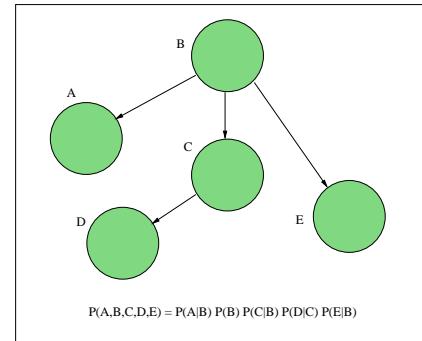


[Devos et al., 2005]

SCCB 2006 ◊ Johan Suykens

75

Kernel design



- Probability product kernel:

$$K(p_1, p_2) = \int_{\mathcal{X}} p_1(x)^\rho p_2(x)^\rho dx$$

- #### - Prior knowledge incorporation

Kernels from graphical models, Bayesian networks, HMMs

Kernels tailored to data types (DNA sequence, text, chemoinformatics)

[Tsuda et al., Bioinformatics 2002; Jebara et al., JMLR 2004, Ralaivola et al., 2005]

SCCB 2006 ◊ Johan Suykens

SCCB 2006 ◊ Johan Suykens

76

Kernels and graphical models

- Probability product kernel [Jebara et al., 2004]

$$K(p_1, p_2) = \int_{\mathcal{X}} p_1(x)^\rho p_2(x)^\rho dx$$

- Case $\rho = 1/2$: Bhattacharyya kernel

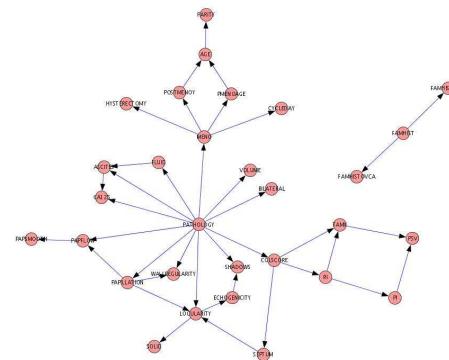
$$K(p_1, p_2) = \int_{\gamma} \sqrt{p_1(x)} \sqrt{p_2(x)} dx$$

(related to Hellinger distance $H(p_1, p_2) = \frac{1}{2} \int_{\mathcal{X}} (\sqrt{p_1(x)} - \sqrt{p_2(x)})^2 dx$ by $H(p_1, p_2) = \sqrt{2 - 2K(p_1, p_2)}$, which is a symmetric approximation to Kullback-Leibler divergence).

- Case $\rho = 1$: expected likelihood kernel

$$K(p_1, p_2) = \int_{\mathcal{X}} p_1(x)p_2(x)dx = \mathbb{E}_{p_1}[p_2(x)] = \mathbb{E}_{p_2}[p_1(x)]$$

Bayesian networks



Example of a Bayesian network, used in the study of ovarian cancer:
Bayesian network approaches enable to incorporate medical expert
knowledge [Fannes et al., 2004]

SCCB 2006 ◊ Johan Suykens

77

SCCB 2006 ◊ Johan Suykens

78

Canonical Correlation Analysis

- CCA analysis has applications e.g. in system identification, signal processing, and recently in bioinformatics and textmining.
- Objective:** find a maximal correlation between the projected variables $z_x = w^T x$ and $z_y = v^T y$ where $x \in \mathbb{R}^{n_x}, y \in \mathbb{R}^{n_y}$ (zero mean).
- Maximize the **correlation coefficient**

$$\max_{w,v} \rho = \frac{\mathcal{E}[z_x z_y]}{\sqrt{\mathcal{E}[z_x z_x] \sqrt{\mathcal{E}[z_y z_y]}}} = \frac{w^T C_{xy} v}{\sqrt{w^T C_{xx} w} \sqrt{v^T C_{yy} v}}$$

with $C_{xx} = \mathcal{E}[xx^T]$, $C_{yy} = \mathcal{E}[yy^T]$, $C_{xy} = \mathcal{E}[xy^T]$. This is formulated as the constrained optimization problem

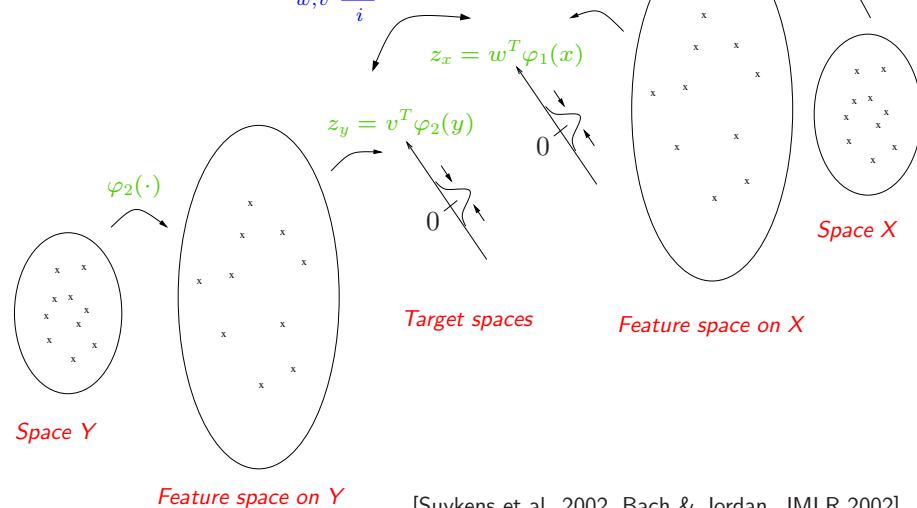
$$\max_{w,v} w^T C_{xy} v \text{ s.t. } w^T C_{xx} w = 1 \text{ and } v^T C_{yy} v = 1$$

which leads to the **generalized eigenvalue problem**

$$C_{xy} v = \eta C_{xx} w, \quad C_{yx} w = \nu C_{yy} v.$$

Kernel CCA

$$\text{Correlation: } \min_{w,v} \sum_i \|z_{x_i} - z_{y_i}\|_2^2$$



LS-SVM formulation to Kernel CCA

- Score variables:** $z_x = w^T (\varphi_1(x) - \hat{\mu}_{\varphi_1})$, $z_y = v^T (\varphi_2(y) - \hat{\mu}_{\varphi_2})$
Feature maps φ_1, φ_2 , kernels $K_1(x_i, x_j) = \varphi_1(x_i)^T \varphi_1(x_j)$, $K_2(y_i, y_j) = \varphi_2(y_i)^T \varphi_2(y_j)$
- Primal problem:** (Kernel PLS case: $\nu_1 = 0, \nu_2 = 0$ [Hoegaerts et al., 2004])

$$\max_{w,v,e,r} \gamma \sum_{i=1}^N e_i r_i - \nu_1 \frac{1}{2} \sum_{i=1}^N e_i^2 - \nu_2 \frac{1}{2} \sum_{i=1}^N r_i^2 - \frac{1}{2} w^T w - \frac{1}{2} v^T v$$

such that $e_i = w^T (\varphi_1(x_i) - \hat{\mu}_{\varphi_1})$, $r_i = v^T (\varphi_2(y_i) - \hat{\mu}_{\varphi_2})$, $\forall i$

with $\hat{\mu}_{\varphi_1} = (1/N) \sum_{i=1}^N \varphi_1(x_i)$, $\hat{\mu}_{\varphi_2} = (1/N) \sum_{i=1}^N \varphi_2(y_i)$.

- Dual problem:** **generalized eigenvalue problem** [Suykens et al. 2002]

$$\begin{bmatrix} 0 & \Omega_{c,2} \\ \Omega_{c,1} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \lambda \begin{bmatrix} \nu_1 \Omega_{c,1} + I & 0 \\ 0 & \nu_2 \Omega_{c,2} + I \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \quad \lambda = 1/\gamma$$

with $\Omega_{c,1ij} = (\varphi_1(x_i) - \hat{\mu}_{\varphi_1})^T (\varphi_1(x_j) - \hat{\mu}_{\varphi_1})$, $\Omega_{c,2ij} = (\varphi_2(y_i) - \hat{\mu}_{\varphi_2})^T (\varphi_2(y_j) - \hat{\mu}_{\varphi_2})$

Obtaining solution from Lagrangian

- Lagrangian $\mathcal{L}(w, v, e, r; \alpha, \beta) = \gamma \sum_{i=1}^N e_i r_i - \nu_1 \frac{1}{2} \sum_{i=1}^N e_i^2 - \nu_2 \frac{1}{2} \sum_{i=1}^N r_i^2 - \frac{1}{2} w^T w - \frac{1}{2} v^T v - \sum_{i=1}^N \alpha_i [e_i - w^T (\varphi_1(x_i) - \hat{\mu}_{\varphi_1})] - \sum_{i=1}^N \beta_i [r_i - v^T (\varphi_2(y_i) - \hat{\mu}_{\varphi_2})]$
- Conditions for optimality (eliminate w, v, e, r)

$$\left\{ \begin{array}{l} \frac{\partial \mathcal{L}}{\partial w} = 0 \rightarrow w = \sum_{i=1}^N \alpha_i (\varphi_1(x_i) - \hat{\mu}_{\varphi_1}) \\ \frac{\partial \mathcal{L}}{\partial v} = 0 \rightarrow v = \sum_{i=1}^N \beta_i (\varphi_2(y_i) - \hat{\mu}_{\varphi_2}) \\ \frac{\partial \mathcal{L}}{\partial e_i} = 0 \rightarrow \gamma v^T (\varphi_2(y_i) - \hat{\mu}_{\varphi_2}) = \nu_1 w^T (\varphi_1(x_i) - \hat{\mu}_{\varphi_1}) + \alpha_i \quad i = 1, \dots, N \\ \frac{\partial \mathcal{L}}{\partial r_i} = 0 \rightarrow \gamma w^T (\varphi_1(x_i) - \hat{\mu}_{\varphi_1}) = \nu_2 v^T (\varphi_2(y_i) - \hat{\mu}_{\varphi_2}) + \beta_i \quad i = 1, \dots, N \\ \frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 \rightarrow e_i = w^T (\varphi_1(x_i) - \hat{\mu}_{\varphi_1}) \quad i = 1, \dots, N \\ \frac{\partial \mathcal{L}}{\partial \beta_i} = 0 \rightarrow r_i = v^T (\varphi_2(y_i) - \hat{\mu}_{\varphi_2}) \quad i = 1, \dots, N \end{array} \right.$$

Kernel CCA applications

- [Vert & Kanehisa, Bioinformatics 2003]:
For kernels related to spaces X and Y

K_1 : graph from gene network
 K_2 : gene expression profiles

Study correlation between gene network and set of profiles

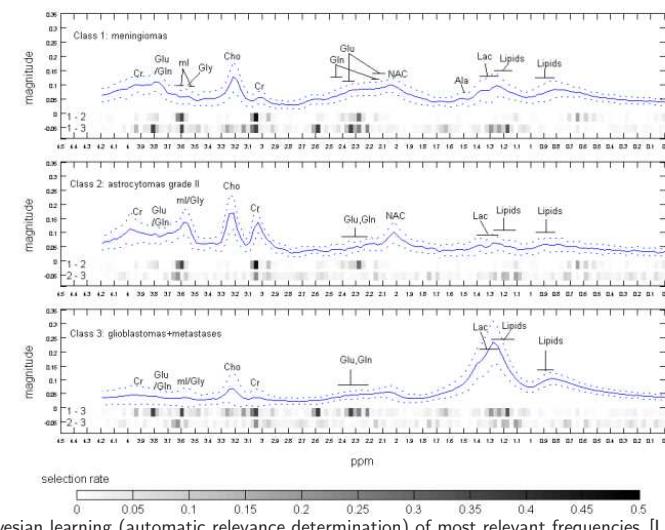
Able to extract biologically relevant expression patterns and pathways with related activity.

- [Yamanishi et al., Bioinformatics 2003]:

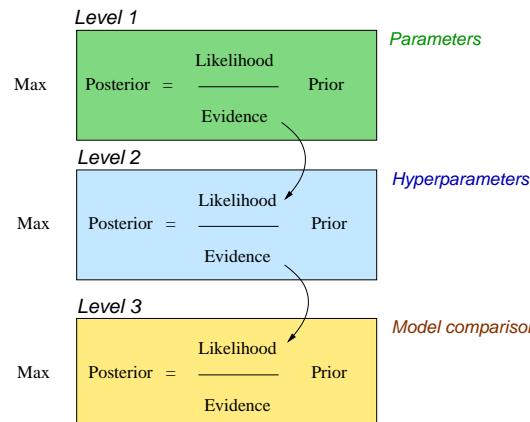
Extract correlated gene clusters from multiple genomic data. Successfully tested on the ability to recognize operons in the *Escherichia coli* genome, from the comparison of three data sets:

- functional relationships between genes in metabolic pathways
- geometrical relationships along the chromosome
- co-expression relationships as observed by gene expression data

Classification of brain tumors using ARD



Bayesian inference



Automatic relevance determination (ARD) [MacKay, 1998]: infer elements of diagonal matrix S in $K(x_i, x_j) = \exp(-(x_i - x_j)^T S(x_i - x_j))$ which indicate how relevant input variables are (but: many local minima, computationally expensive).

Additive regularization trade-off

- Traditional Tikhonov regularization scheme:

$$\min_{w,e} w^T w + \gamma \sum_i e_i^2 \text{ s.t. } e_i = y_i - w^T \varphi(x_i), \quad \forall i = 1, \dots, N$$

Training solution for fixed value of γ :

$$(K + I/\gamma)\alpha = y$$

→ Selection of γ via validation set: **non-convex** problem

- Additive regularization trade-off [Pelckmans et al., 2005]:

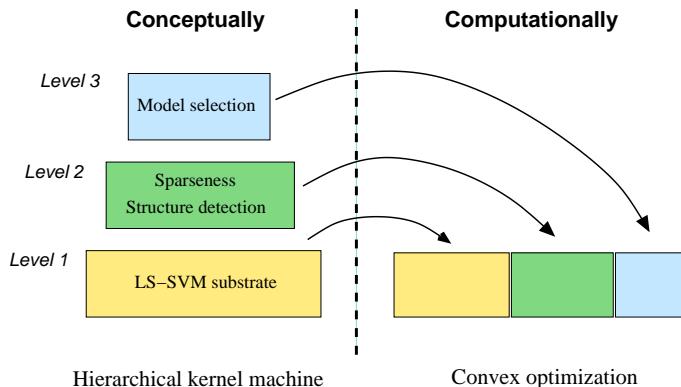
$$\min_{w,e} w^T w + \sum_i (e_i - c_i)^2 \text{ s.t. } e_i = y_i - w^T \varphi(x_i), \quad \forall i = 1, \dots, N$$

Training solution for fixed value of $c = [c_1; \dots; c_N]$:

$$(K + I)\alpha = y - c$$

→ Selection of c via validation set: can be **convex** problem

Hierarchical Kernel Machines



Hierarchical modelling approach leading to convex optimization problem
 Computationally fusing training, hyperparameter and model selection
 Optimization modelling: sparseness, input/structure selection, stability ...

[Pelckmans et al., Machine Learning 2006]

SCCB 2006 ◊ Johan Suykens

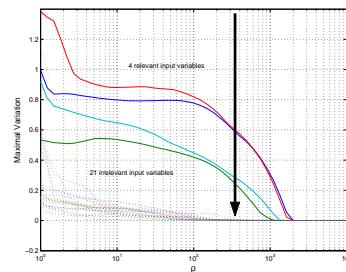
87

Additive models and structure detection

- **Additive models:** $\hat{y}(x) = \sum_{p=1}^P w^{(p)T} \varphi^{(p)}(x^{(p)})$ with $x^{(p)}$ the p -th input variable and a feature map $\varphi^{(p)}$ for each variable. This leads to the kernel $K(x_i, x_j) = \sum_{p=1}^P K^{(p)}(x_i^{(p)}, x_j^{(p)})$.
- **Structure detection** [Pelckmans et al., 2005]:

$$\begin{aligned} & \min_{w, e, t} \rho \sum_{p=1}^P t_p + \sum_{p=1}^P w^{(p)T} w^{(p)} + \gamma \sum_{i=1}^N e_i^2 \\ & \text{s.t. } \begin{cases} y_i = \sum_{p=1}^P w^{(p)T} \varphi^{(p)}(x_i^{(p)}) + e_i, & \forall i = 1, \dots, N \\ -t_p \leq w^{(p)T} \varphi^{(p)}(x_i^{(p)}) \leq t_p, & \forall i = 1, \dots, N, \forall p = 1, \dots, P \end{cases} \end{aligned}$$

Study how the solution with maximal variation varies for different values of ρ



SCCB 2006 ◊ Johan Suykens

89

Issues about sparseness

- Sparse approximation (zeros in solution vector) - two main approaches in general:
 1. by choice of loss function (e.g. epsilon-insensitive loss function)
 2. by choice of the regularization term (e.g. 1-norm instead of 2-norm)
- For linear models (or parameterized models): both options possible
- For support vector machines: rely on 2-norm for regularization term (for kernel based model representation in dual) and w can be infinite dimensional (e.g. in RBF kernel case).
- Interpretability helps with additive models [Hastie & Tibshirani, 1986] and componentwise models (also suitable in high dimensions)

SCCB 2006 ◊ Johan Suykens

88

Incorporation of prior knowledge (1)

- Support vector machine formulations allow to incorporate additional constraints that express prior knowledge about the problem, e.g. monotonicity, symmetry, positivity, ...
- Especially, LS-SVM as simple core models for which one can take into account additional regularization terms and/or constraints. Systematic and straightforward (dual) solution from Lagrangian.
- Large potential of convex optimization techniques [Boyd & Vandenberghe, 2004]

SCCB 2006 ◊ Johan Suykens

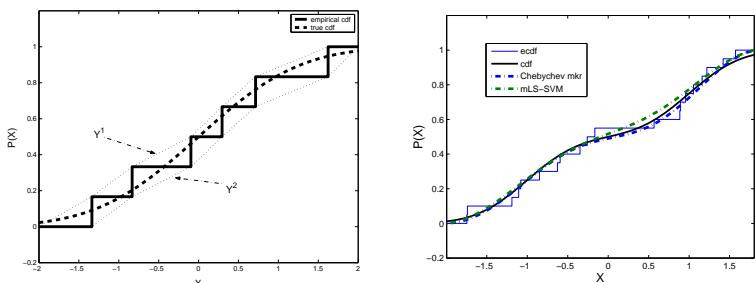
90

Incorporation of prior knowledge (2)

- Example: LS-SVM regression with monotonicity constraint

$$\min_{w,b,e} \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^N e_i^2 \quad \text{s.t.} \quad \begin{cases} y_i = w^T \varphi(x_i) + b + e_i, \forall i = 1, \dots, N \\ w^T \varphi(x_i) \leq w^T \varphi(x_{i+1}), \forall i = 1, \dots, N-1 \end{cases}$$

- Application: estimation of cdf [Pelckmans et al., 2005]



Acknowledgements (1)

- K.U. Leuven, ESAT-SCD: research teams SMC, BIOI, Biomed
Prof. B. De Moor, Prof. S. Van Huffel, Prof. K. Marchal, Prof. Y. Moreau, Prof. J. Vandewalle

Current and former postdocs and PhD candidates:

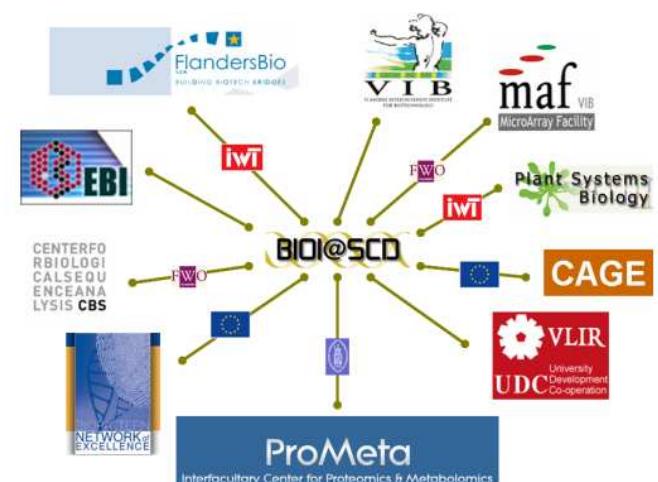
C. Alzate, Dr. L. Ameye, Dr. T. De Bie, Dr. J. De Brabanter, Dr. A. Devos, Dr. M. Espinoza, Dr. F. De Smet, O. Gevaert, Dr. I. Goethals, Dr. B. Hamers, F. Janssens, Dr. L. Hoegaerts, P. Karsmakers, Dr. G. Lanckriet, Dr. C. Lu, Dr. L. Lukas, J. Luts, F. Ojeda, Dr. K. Pelckmans, Dr. N. Pochet, B. Van Calster, R. Van de Plas, Dr. T. Van Gestel, V. Van Belle, Dr. L. Vanhamme

- Many people for joint work, discussions, invitations, joint organization of meetings.

Acknowledgements (2)

- Support from GOA-Ambiorics (Algorithms for Medical and Biological Research, Integration, Computation and Software), COE Optimization in Engineering, COE Symbiosys, IAP V, FWO projects, IWT.
- Biomed MRS/MRSI research in collaboration with biomedical NMR unit, Dept. of Radiology, Univ. Hospitals Leuven, Belgium, partners of EU projects INTERPRET (IST-1999-10310), eTUMOUR (FP6-2002-LIFESCIHEALTH 503094), BIOPATTERN (FP6-2002-IST 508803), HEALTHagents (IST-2004-27214)

Acknowledgements (3)



References: books

- Boyd S., Vandenberghe L., *Convex Optimization*, Cambridge University Press, 2004.
- Chapelle O., Schölkopf B., Zien A. (Eds.), *Semi-Supervised Learning*, MIT Press, Cambridge, MA, (in press) 2006.
- Cristianini N., Shawe-Taylor J., *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.
- Rasmussen C.E., Williams C.K.I., *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA 2006.
- Schölkopf B., Smola A., *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.
- Schölkopf B., Tsuda K., Vert J.P. (Eds.) *Kernel Methods in Computational Biology* 400, MIT Press, Cambridge, MA (2004)
- Shawe-Taylor J., Cristianini N., *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.
- Suykens J.A.K., Van Gestel T., De Brabanter J., De Moor B., Vandewalle J., *Least Squares Support Vector Machines*, World Scientific, Singapore, 2002.
- Suykens J.A.K., Horvath G., Basu S., Micchelli C., Vandewalle J. (Eds.), *Advances in Learning Theory : Methods, Models and Applications*, vol. 190 of NATO-ASI Series III : Computer and Systems Sciences, IOS Press (Amsterdam, The Netherlands) 2003.
- Vapnik V., *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.
- Wahba G., *Spline Models for Observational Data*, Series in Applied Mathematics, 59, SIAM, Philadelphia, 1990.

Related references - methods

- Alzate C., Suykens J. A. K., "A Weighted Kernel PCA Formulation with Out-of-Sample Extensions for Spectral Clustering Methods", WCCI-IJCNN 2006, Vancouver, 138-144.
- Bach F.R., Jordan M.I., "Kernel independent component analysis", *Journal of Machine Learning Research*, 3, 1-48, 2002.
- Belkin M., Niyogi P., "Semi-supervised learning on Riemannian manifolds", *Machine Learning*, Vol. 56, pp. 209-239, 2004.
- Burges C.J.C., "A tutorial on support vector machines for pattern recognition", *Knowledge Discovery and Data Mining*, 2(2), 121-167, 1998.
- Cawley G.C., Talbot N.L.C., "Fast exact leave-one-out cross-validation of sparse least-squares support vector machines", *Neural Networks*, Vol. 17, 10, pp. 1467-1475, 2004.
- Cawley G.C., "Leave-one-out Cross-validation Based Model Selection Criteria for Weighted LS-SVMs", WCCI-IJCNN 2006 Vancouver.
- Chung F.R.K. , "Spectral graph theory", Regional Conference Series in Mathematics 92, Amer. Math. Soc., Providence, 1997.
- Cortes C., Vapnik V., "Support vector networks", *Machine Learning*, 20, 273-297, 1995.
- Cucker F., Smale S., "On the mathematical foundations of learning theory", *Bulletin of the AMS*, 39, 1-49, 2002
- Debruyne M., Christmann A., Hubert M., Suykens J.A.K., "Robustness and stability of reweighted kernel based regression", Internal Report 06-150, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2006.
- Espinoza M., Suykens J.A.K., De Moor B., "Kernel Based Partially Linear Models and Nonlinear Identification", *IEEE Transactions on Automatic control, special issue (System identification : linear vs nonlinear)*, vol. 50, no. 10, Oct. 2005, pp. 1509- 1519.

- Espinoza M., Suykens J.A.K., De Moor B., "LS-SVM Regression with Autocorrelated Errors", in *Proc. of the 14th IFAC Symposium on System Identification (SYSID)*, Newcastle, Australia, Mar. 2006, pp. 582-587.
- Evgeniou T., Pontil M., Poggio T., "Regularization networks and support vector machines", *Advances in Computational Mathematics*, 13(1): 1-50, 2000.
- Girolami M., "Orthogonal series density estimation and the kernel eigenvalue problem", *Neural Computation*, 14(3), 669-688, 2002.
- Girosi F., "An equivalence between sparse approximation and support vector machines", *Neural Computation*, 10(6), 1455-1480, 1998.
- Hastie T., Tibshirani R., "Generalized Additive Models (with discussion)", *Statistical Science*, Vol 1, No 3, 297-318, 1986.
- Hoegaerts L., Suykens J.A.K., Vandewalle J., De Moor B., "Primal space sparse kernel partial least squares regression for large scale problems", IJCNN 2004, Hungary, Budapest, Jul. 2004, pp. 561-566.
- Hoegaerts L., Suykens J.A.K., Vandewalle J., De Moor B., "Subset based least squares subspace regression in RKHS", *Neurocomputing*, vol. 63, Jan. 2005, pp. 293-323.
- Jebara T., Kondor R., Howard A., "Probability Product Kernels", *Journal of Machine Learning Research*, 5(Jul):819-844, 2004.
- Kondor R., Lafferty J., "Diffusion Kernels on Graphs and Other Discrete Input Spaces". ICML 2002.
- Kwok J.T., "The evidence framework applied to support vector machines", *IEEE Transactions on Neural Networks*, 10, 1018-1031, 2000.
- Lanckriet, G.R.G., Cristianini, N., Bartlett, P., El Ghaoui, L., Jordan, M.I., "Learning the Kernel Matrix with Semidefinite Programming", *Journal of Machine Learning Research*, 5, 27-72, 2004.
- Lin C.-J., "On the convergence of the decomposition method for support vector machines", *IEEE Transactions on Neural Networks*. 12, 1288-1298, 2001.

- MacKay D.J.C., "Bayesian interpolation", *Neural Computation*, 4(3), 415-447, 1992.
- MacKay D.J.C., "Introduction to Gaussian processes", In C.M. Bishop, editor, *Neural Networks and Machine Learning*, volume 168 of NATO ASI Series, pages 133-165. Springer, Berlin, 1998.
- Mercer J., "Functions of positive and negative type and their connection with the theory of integral equations", *Philos. Trans. Roy. Soc. London*, 209, 415-446, 1909.
- Mika S., Rätsch G., Weston J., Schölkopf B., Müller K.-R., "Fisher discriminant analysis with kernels", In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, 41-48. IEEE, 1999.
- Müller K.R., Mika S., Rätsch G., Tsuda K., Schölkopf B., "An introduction to kernel-based learning algorithms", *IEEE Transactions on Neural Networks*, 2001. 12(2): 181-201, 2001.
- Ng A.Y., Jordan M.I., Weiss Y., "On spectral clustering: Analysis and an algorithm", In T. Dietterich, S. Becker and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems (NIPS) 14*, 2002.
- Pelckmans K., Suykens J.A.K., De Moor B., "Building Sparse Representations and Structure Determination on LS-SVM Substrates", *Neurocomputing*, vol. 64, Mar. 2005, pp. 137-159.
- Pelckmans K., Suykens J.A.K., De Moor B., "Additive regularization trade-off: fusion of training and validation levels in kernel methods", *Machine Learning*, vol. 62, no. 3, Mar. 2006, pp. 217-252.
- Pelckmans K., De Brabanter J., Suykens J.A.K., De Moor B., "Handling Missing Values in Support Vector Machine Classifiers", *Neural Networks*, vol. 18, 2005, pp. 684-692.
- Pelckmans K., Espinoza M., De Brabanter J., Suykens J.A.K., De Moor B., "Primal-Dual Monotone Kernel Regression", *Neural processing letters*, vol. 22, no. 2, Oct. 2005, pp. 171-182..
- Pelckmans K., De Brabanter J., Suykens J.A.K., De Moor B., "Convex Clustering Shrinkage", in Workshop on Statistics and Optimization of Clustering Workshop (PASCAL), London, U.K., Jul. 2005
- Perez-Cruz F., Bousono-Calzon C., Artes-Rodriguez A., "Convergence of the IRWLS Procedure to the Support Vector Machine Solution", *Neural Computation*, 17: 7-18, 2005.

- Platt J., "Fast training of support vector machines using sequential minimal optimization", In Schölkopf B., Burges C.J.C., Smola A.J. (Eds.) *Advances in Kernel methods - Support Vector Learning*, 185–208, MIT Press, 1999.
- Poggio T., Girosi F., "Networks for approximation and learning", *Proceedings of the IEEE*, **78**(9), 1481–1497, 1990.
- Poggio T., Rifkin R., Mukherjee S., Niyogi P., "General conditions for predictivity in learning theory", *Nature*, **428** (6981): 419–422, 2004.
- Principe J., Fisher III, Xu D., "Information theoretic learning", in S. Haykin (Ed.), *Unsupervised Adaptive Filtering*, John Wiley & Sons, New York, 2000.
- Rosipal R., Trejo L.J., "Kernel partial least squares regression in reproducing kernel Hilbert space", *Journal of Machine Learning Research*, **2**, 97–123, 2001.
- Saunders C., Gammerman A., Vovk V., "Ridge regression learning algorithm in dual variables", *Proc. of the 15th Int. Conf. on Machine Learning (ICML-98)*, Madison-Wisconsin, 515–521, 1998.
- Schölkopf B., Smola A., Müller K.-R., "Nonlinear component analysis as a kernel eigenvalue problem", *Neural Computation*, **10**, 1299–1319, 1998.
- Schölkopf B., Mika S., Burges C., Knirsch P., Müller K.-R., Rätsch G., Smola A., "Input space vs. feature space in kernel-based methods", *IEEE Transactions on Neural Networks*, **10**(5), 1000–1017, 1999.
- Shi J., Malik J., "Normalized Cuts and Image Segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), 888–905, 2000.
- Suykens J.A.K., Vandewalle J., "Least squares support vector machine classifiers", *Neural Processing Letters*, vol. 9, no. 3, Jun. 1999, pp. 293–300.
- Suykens J.A.K., De Brabanter J., Lukas L., Vandewalle J., "Weighted least squares support vector machines : robustness and sparse approximation", *Neurocomputing, Special issue on fundamental and information processing aspects of neurocomputing*, vol. 48, no. 1-4, Oct. 2002, pp. 85–105.

Related references - applications

- Aerts S., Lambrechts D., Maity S., Van Loo P., Coessens B., De Smet F., Tranchevent L.C., De Moor B., Marynen P., Hassan B., Carmeliet P., Moreau Y., "Gene prioritization through genomic data fusion", *Nature Biotechnology*, **24** (5): 537–544, 2006.
- Antal P., Fannes G., Timmerman D., Moreau Y., De Moor B., "Using literature and data to learn Bayesian Networks as clinical models of ovarian tumors", *Artificial Intelligence in Medicine*, vol. 30, 2004, pp. 257–281.
- Brown M., Grundy W., Lin D., Cristianini N., Sugnet C., Furey T., Ares M., Haussler D., "Knowledge-based analysis of microarray gene expression data using support vector machines", *Proceedings of the National Academy of Science*, **97**(1), 262–267, 2000.
- Devos A., Lukas L., Suykens J.A.K., Vanhamme L., Tate A.R., Howe F.A., Majos C., Moreno-Torres A., Van der Graaf M., Arus C., Van Huffel S., "Classification of brain tumours using short echo time 1H MRS spectra", *Journal of Magnetic Resonance*, vol. 170 , no. 1, Sep. 2004, pp. 164–175.
- Devos A., Simonetti A.W., van der Graaf M., Lukas L., Suykens J.A.K., Vanhamme L., Buydens L.M.C., Heerschap A., Van Huffel S. "The use of multivariate MR imaging intensities versus metabolic data from MR spectroscopic imaging for brain tumour classification", *Journal of Magnetic Resonance*, **173** (2): 218–228 April 2005.
- Guyon I., Weston J., Barnhill S., Vapnik V., "Gene selection for cancer classification using support vector machines", *Machine Learning*, **46**, 389–422, 2002.
- Lanckriet, G.R.G., De Bie, T., Cristianini, N. , Jordan, M.I., Noble, W.S., "A statistical framework for genomic data fusion", *Bioinformatics*, **20**, 2626–2635, 2004.
- Lu C., Van Gestel T., Suykens J.A.K., Van Huffel S., Vergote I., Timmerman D., "Preoperative prediction of malignancy of ovarian tumor using least squares support vector machines", *Artificial Intelligence in Medicine*, vol. 28, no. 3, Jul. 2003, pp. 281–306.

- Suykens J.A.K., Van Gestel T., Vandewalle J., De Moor B., "A support vector machine formulation to PCA analysis and its kernel version", *IEEE Transactions on Neural Networks*, vol. 14, no. 2, Mar. 2003, pp. 447–450.
- Van Gestel T., Suykens J.A.K., Baesens B., Viaene S., Vanthienen J., Dedene G., De Moor B., Vandewalle J., "Benchmarking Least Squares Support Vector Machine Classifiers", *Machine Learning*, vol. 54, no. 1, Jan. 2004, pp. 5–32.
- Van Gestel T., Suykens J.A.K., Lanckriet G., Lambrechts A., De Moor B., Vandewalle J., "Bayesian Framework for Least Squares Support Vector Machine Classifiers, Gaussian Processes and Kernel Fisher Discriminant Analysis", *Neural Computation*, vol. 15, no. 5, May 2002, pp. 1115–1148.
- Williams C.K.I., Rasmussen C.E., "Gaussian processes for regression", In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems* **8**, 514–520. MIT Press, 1996.
- Williams C.K.I., Seeger M., "Using the Nyström method to speed up kernel machines", In T.K. Leen, T.G. Dietterich, and V. Tresp (Eds.), *Advances in neural information processing systems*, **13**, 682–688, MIT Press, 2001.
- Zanni L., Serafini T., Zanghirati G., "Parallel Software for Training Large Scale Support Vector Machines on Multiprocessor Systems", *Journal of Machine Learning Research*, **7**:1467–1492, 2006.

- Luts J., Heerschap A., Suykens J.A.K., Van Huffel S., "A combined MRI and MRSI based Multiclass System for Brain Tumour Recognition using LS-SVMs with Class Probabilities and Feature Selection", Internal Report 06-143, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2006.
- Pochet N., De Smet F., Suykens J.A.K., De Moor B., "Systematic benchmarking of microarray data classification : assessing the role of nonlinearity and dimensionality reduction", *Bioinformatics*, vol. 20, no. 17, Nov. 2004, pp. 3185–3195.
- Pochet N.L.M.M., Janssens F.A.L., De Smet F., Marchal K., Suykens J.A.K., De Moor B.L.R., "M@CBETH: a microarray classification benchmarking tool", *Bioinformatics*, vol. 21, no. 14, Jul. 2005, pp. 3185–3186.
- Ralaivola L., Swamidass S., Saigo H., Baldi P., "Graph Kernels for Chemical Informatics", *Neural Networks*, **18**(8): 1093–1110, 2005.
- Tyers M., Mann M., "From genomics to proteomics", *Nature*, Vol. 422, 193–197, 2003.
- Tsuda K., Kin T., Asai K. "Marginalized kernels for biological sequences", *Bioinformatics*, **18**(Suppl.1): S268–S275, 2002.
- Tsuda K., Shin H.J., Schölkopf B., "Fast protein classification with multiple networks", *Bioinformatics (ECCB'05)*, **21**(Suppl. 2):ii59–ii65, 2005.
- Vert J.-P., Kanehisa M., "Extracting active pathways from gene expression data", *Bioinformatics*, vol. 19, p. 238ii–244ii, 2003.
- Yamanishi Y., Vert J.-P., Nakaya A., Kanehisa M., "Extraction of Correlated Gene Clusters from Multiple Genomic Data by Generalized Kernel Canonical Correlation Analysis", *Bioinformatics*, vol. 19, p. 323i–330i, 2003. (Proceedings of ISMB 2003).