Linear Algebra in and for Least Squares Support Vector Machines

bart.demoor@kuleuven.be
www.bartdemoor.be

Katholieke Universiteit Leuven
Department of Electrical Engineering
ESAT-STADIUS

Outline

Outline

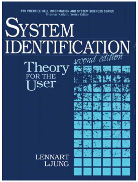
- Big Data: Volume, Velocity, Variety, . . .
- Dealing with high dimensional input spaces
- Need for powerful black box modeling techniques
- Avoid pitfalls of nonlinear optimization (convergence issues, local minima,. . .)
- Preferable: (numerical) linear algebra, convex optimization
- Algorithms for (un-)supervised function regression and estimation, (predictive) modeling, clustering and classification, data dimensionality reduction, correlation analysis (spatial-temporal modeling), feature selection, (early - intermediate - late) data fusion, ranking, outlier and fraud detection, decision support systems (process industry 4.0, digital health, . . .)
- . . .

	Linear Algebra	LS-SVM/Kernel
Supervised	Least squares Classification	Function Estimation Kernel Classification
Unsupervised	SVD - PCA Angles - CCA	Kernel PCA Kernel CCA

Outline

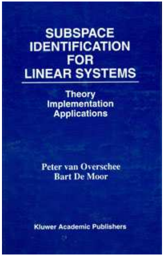
System Identification: PEM

- LTI models
- Non-convex optimization
- Considered 'solved' early nineties



Linear Algebra approach

⇒ **Large block Hankel data matrices; SVD; Orthogonal and oblique projections; Least squares**
⇒ **Subspace methods**



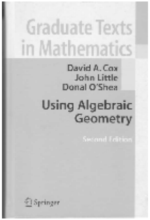
Multivariate polynomial optimization problems

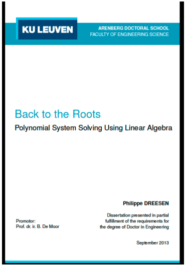
- Multivariate polynomial object function + constraints
- Non-convex optimization
- Computer Algebra, Homotopy methods, Numerical Optimization

Linear Algebra approach

⇒ **Macaulay matrix; SVD; Kernel; Realization theory**

⇒ **Smallest eigenvalue of large matrix**





KATHOLIEKE UNIVERSITEIT LEUVEN ESAT SCD

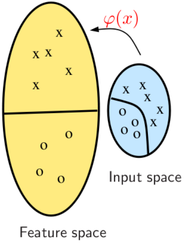
Nonlinear regression, modelling and clustering

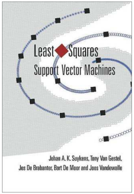
- Most regression, modelling and clustering problems are nonlinear when formulated in the input data space
- This requires nonlinear nonconvex optimization algorithms

Linear Algebra approach

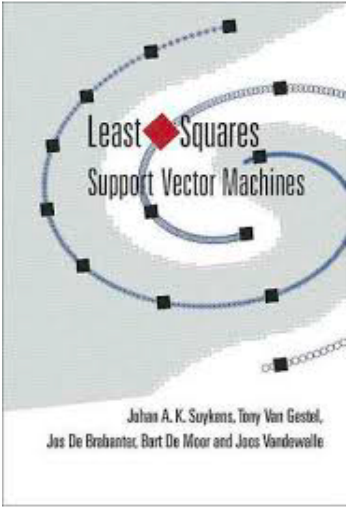
⇒ **Least Squares Support Vector Machines**

- 'Kernel trick' = projection of input data to a high-dimensional feature space
- Regression, modelling, clustering problem becomes a large scale linear algebra problem (set of linear equations, eigenvalue problem)





Least Squares Support Vector Machines. . .



	Linear Algebra	**LS-SVM/Kernel**
Supervised	**Least squares** Classification	**Function Estimation** Kernel Classification
Unsupervised	SVD - PCA Angles - CCA	Kernel PCA Kernel CCA

Outline

$$X.w = y$$

Consistent $\iff \mathrm{rank}(X) = \mathrm{rank}(X\ y)$

Then estimate w unique iff X of full column rank

Inconsistent if $\mathrm{rank}(X\ y) = \mathrm{rank}(X) + 1$

Find ‘best’ linear combination of columns of X to approximate $y \implies$ **Least Squares**

$$\min_w \|y - Xw\|_2^2 =$$
$$\min_w \; w^T X^T X w - 2y^T X w + y^T y$$

Derivatives w.r.t. $w \Longrightarrow$ **normal equations**

$$X^T X w = X^T y$$

If X full column rank: $w = (X^T X)^{-1} X^T y$

Equivalently: Call $e = y - Xw$

$\min_{e,w} \|e\|_2^2 = e^T e$ subject to $y - Xw - e = 0$

Lagrangean $\mathcal{L}(e, w, l) = \frac{1}{2} e^T e + l^T (y - Xw - e)$

$$\frac{\partial \mathcal{L}}{\partial e} = 0 = e - l \implies e = l$$
$$\frac{\partial \mathcal{L}}{\partial w} = 0 = X^T l \implies X^T e = 0$$
$$\frac{\partial \mathcal{L}}{\partial l} = 0 = y - Xw - e \implies \boxed{X^T X w = X^T y}$$

Consider

$$\min_{e,w}\ \frac{1}{2}e^T V^{-1} e + \frac{1}{2} w^T W^{-1} w$$

subject to

$$y = Xw + e$$

This is maximum likelihood/Bayesian with priors

$$e \sim \mathcal{N}(0, V) \text{ and } w \sim \mathcal{N}(0, W)\ .$$

Lagrangean

$$\mathcal{L}(w, l, e) = \frac{1}{2}e^T V^{-1} e + \frac{1}{2} w^T W^{-1} w - l^T (y - Xw - e)$$

$$\frac{\partial \mathcal{L}}{\partial w} = 0 = W^{-1}w + X^T l$$
$$\frac{\partial \mathcal{L}}{\partial l} = 0 = y - Xw - e$$
$$\frac{\partial \mathcal{L}}{\partial e} = 0 = V^{-1}e + l$$

Hence $\boxed{\begin{pmatrix} 0 & X & I \\ X^T & W^{-1} & 0 \\ I & 0 & V^{-1} \end{pmatrix} \begin{pmatrix} l \\ w \\ e \end{pmatrix} = \begin{pmatrix} y \\ 0 \\ 0 \end{pmatrix}}$

Karush-Kuhn-Tucker equations

Let $X \in \mathbf{R}^{p\times q}$. Eliminate $e = y - Xw$

$$V^{-1}y - V^{-1}Xw + l = 0$$
$$W^{-1}w + X^T l = 0$$

Primal: Eliminate l

$$x = (X^T V^{-1} X + W^{-1})^{-1} X^T V^{-1} y$$

$\rightarrow$ 'small' $q \times q$ inverse

Dual: Eliminate w

$$l = -(V + XWX^T)^{-1} y$$

$\rightarrow$ 'large' $p \times p$ inverse

Given $X \in \mathbf{R}^{N \times q}, y \in \mathbf{R}^N$ with i-th row x_i.
Consider a nonlinear vector function $\varphi(x_i) \in \mathbf{R}^{1 \times q}$ and the constrained least squares optimization problem:

$$\min_{w,b,e} \frac{1}{2} w^T w + \frac{\gamma}{2} e^T e$$

subject to

$$y = \begin{pmatrix} \varphi(x_1) \\ \vdots \\ \varphi(x_N) \end{pmatrix} w + e = X_\varphi w + e$$

Lagrangean

$$\mathcal{L}(w, e, l) = \frac{1}{2} w^T w + \frac{\gamma}{2} e^T e + l^T (y - X_\varphi w - e)$$

Eliminate e and w to find

$$l = (X_\varphi X_\varphi^T + \frac{1}{\gamma} I_N)^{-1} y$$

Call $(X_\varphi X_\varphi^T + \frac{1}{\gamma} I_N)$ the kernel $K(.,.)$ with element i, j: $K(x_i, x_j) = \varphi(x_i).\varphi^T(x_j)$, an 'inner product'.

Then, obviously

$$\boxed{\text{y(x)} = \sum_{i=1}^{N} K(x_i, x) l_i}$$

LS-SVM regression: dual problem

Model:

$$\hat{y} = \sum_i \alpha_i K(x_i, x) + b$$

where α, b follows from

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

where

$$\Omega_{ij} = \varphi(x_i)^T \varphi(x_j) = K(x_i, x_j)$$

for $i, j = 1, ..., N$ and $y = [y_1; ...; y_N]$.

Observations:

- Kernel: $N \times N$, symmetric positive definite matrix
- q can be large (possibly ∞)
- $\varphi(x_i)$ nonlinearly maps data row x_i into a high-(possibly ∞-)dimensional space.
- Not needed to know $\varphi(.)$ explicitly. In machine learning, we fix a symmetric continuous kernel that satisfies Mercer's condition:

$$\int K(x,z)g(x)g(z)dxdz \geq 0 \ ,$$

  for any square integrable function $g(x)$. Then $K(x,z)$ separates: $\exists$ Hilbert space $\mathcal{H}$, $\exists$ map $\phi(.)$ and $\exists \ \lambda_i > 0$ such that

$$K(x,z) = \sum \lambda_i \phi(x)\phi(z) \ .$$

- **Kernel trick:** Work out 'dual formulation' with Lagrange multipliers; generate 'long' (∞) inner products with $\varphi(.)$;

Kernels:

Mathematical form: linear, polynomial, radial basis function, splines, wavelets, string kernel, kernels from graphical models, Fisher kernels, graph kernels, data fusion kernels, spike kernels, ...

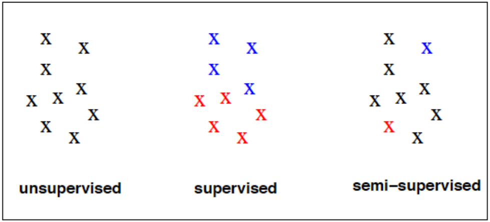
Application inspired: Text mining, bioinformatics, images, ...

$$K(x, x_i) = x_i^T x \text{ (linear SVM)}$$
$$K(x, x_i) = (x_i^T x + \tau)^d \text{ (polynomial SVM of degree } d\text{)}, \tau \geq 0$$
$$K(x, x_i) = \exp(-\|x - x_i\|_2^2/\sigma^2) \text{ (RBF kernel)}$$
$$K(x, x_i) = \tanh(\kappa\, x_i^T x + \theta) \text{ (MLP kernel)}$$

	Linear Algebra	LS-SVM/Kernel
Supervised	Least squares **Classification**	Function Estimation **Kernel Classification**
Unsupervised	SVD - PCA Angles - CCA	Kernel PCA Kernel CCA

Outline

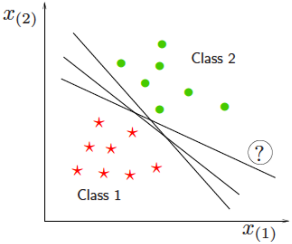
Learning: unsupervised, supervised, semi-supervised



Given data can be labeled, unlabeled or partially labeled
(clustering = unsupervised, classification = supervised)

Training set $\{(x_i, y_i)\}_{i=1}^N$:
input data $x_i \in \mathbb{R}^d$
class labels $y_i \in \{-1, +1\}$

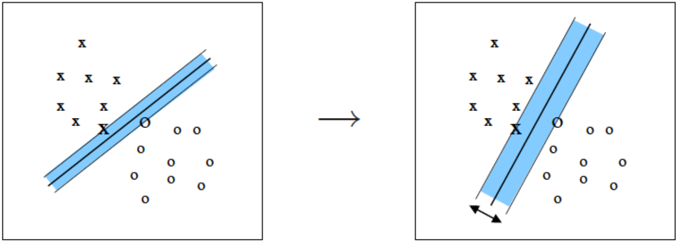
Classifier: $\hat{y} = \text{sign}[w^T x + b]$



Requirement that all training data are correctly classified:

$$\begin{aligned} & w^T x_i + b \geq +1, \quad \text{if} \quad y_i = +1 \\ & w^T x_i + b \leq -1, \quad \text{if} \quad y_i = -1 \\ \Leftrightarrow \quad & y_i [w^T x_i + b] \geq 1, \quad \forall i \end{aligned}$$

SVM: maximize the margin



Margin $= \frac{2}{\|w\|}$

$$\min_{w,b} \quad \frac{1}{2} w^T w$$
$$\text{subject to} \quad y_i[w^T x_i + b] \geq 1 \quad , \ i = 1, ..., N$$

- Preserve support vector machine [Vapnik, 1995] methodology, but simplify via least squares and equality constraints [Suykens, 1999]
- **Primal problem:**

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

- **Dual problem:**

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

where $\Omega_{ij} = y_i y_j \, \varphi(x_i)^T \varphi(x_j) = y_i y_j \, K(x_i, x_j)$ and $y = [y_1; \ldots; y_N]$.

- LS-SVM classifiers perform very well on 20 UCI data sets [Van Gestel et al., ML 2004] Winning results in competition WCCI 2006 by [Cawley, 2006]

- 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.

- Conditions for optimality:

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

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

Outline

	Linear Algebra	LS-SVM/Kernel
Supervised	Least squares Classification	Function Estimation Kernel Classification
Unsupervised	**SVD - PCA** Angles - CCA	**Kernel PCA** Kernel CCA

Given data matrix X. Find vectors of maximum variance:

$$\max_{w,e} e^T e \ ,$$

subject to

$$e = Xw \ , \ w^T w = 1 \ .$$

Lagrangean:

$$\mathcal{L}(w, e, \lambda) = \frac{1}{2} e^T e - l^T (e - Xw) + \lambda (1 - w^T w)$$

giving

$$\begin{aligned} 0 &= e - l \\ 0 &= X^T l - 2w\lambda \\ 0 &= e - Xw \\ 1 &= w^T w \end{aligned}$$

Hence

$$l = Xw \ , \ X^T l = w(2\lambda) \ , l^T l = 2\lambda \ .$$

Call $v = l/\sqrt{2\lambda}$ and $\sigma = \sqrt{2\lambda}$, then

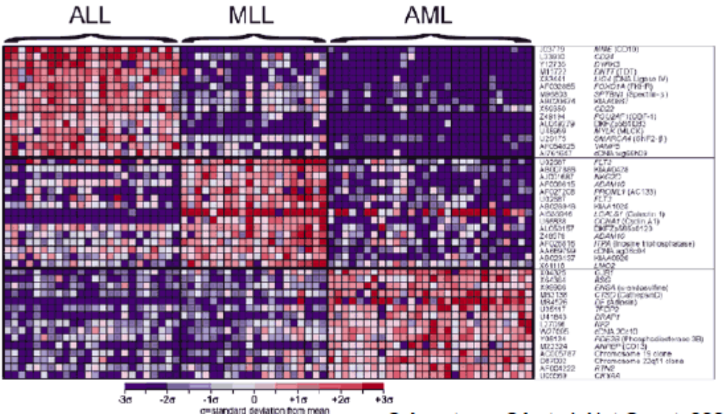
$$\begin{aligned} Xw = v\sigma \quad &, \quad w^T w = 1 \\ X^T v = w\sigma \quad &, \quad v^T v = 1 \end{aligned}$$

SVD !! So, the left singular vectors of X are Lagrange multipliers in a PCA problem.

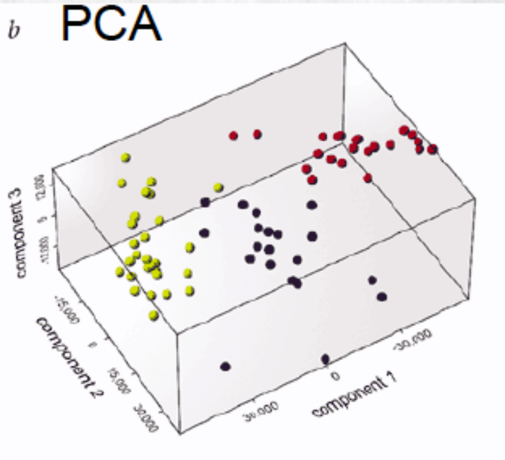
Example: 12 600 genes
72 patients:

- 28 Acute Lymphoblastic Leukemia (ALL)
- 24 Acute Myeloid Leukemia (AML)
- 20 Mixed Linkage Leukemia (MLL)

PCA - SVD



PCA - SVD


b
PCA
component 3
component 2
component 1
0
30,000
-30,000
15,000
-15,000

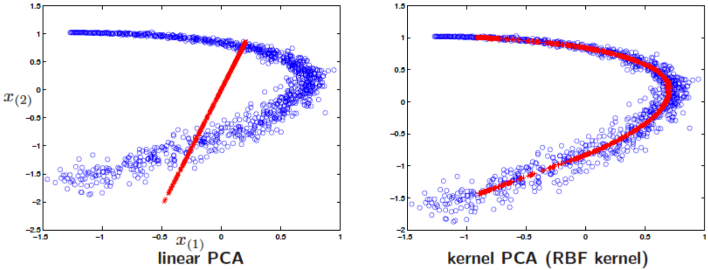
- Primal problem:

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

- Dual problem = kernel PCA :

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

with $\Omega_{c,ij} = (\varphi(x_i) - \hat{\mu}_\varphi)^T (\varphi(x_j) - \hat{\mu}_\varphi)$ in *centered kernel matrix.*



Kernel PCA [Schölkopf et al., 1998]:

$$\text{eigenvalue decomposition of} \quad \begin{bmatrix} K(x_1, x_1) & \dots & K(x_1, x_N) \\ \vdots & & \vdots \\ K(x_N, x_1) & \dots & K(x_N, x_N) \end{bmatrix}$$

Outline

	Linear Algebra	LS-SVM/Kernel
Supervised	Least squares Classification	Function Estimation Kernel Classification
Unsupervised	SVD - PCA **Angles - CCA**	Kernel PCA **Kernel CCA**

Given two data matrices X, Y. Find directions in the column space of X, resp. Y that maximally correlate:

$$\min_{e,f,v,w} \frac{1}{2}\|e - f\|_2^2 \ ,$$

subject to

$$e = Xv, f = Yw, e^T e = 1, f^T f = 1 \ .$$

Notice that

$$\|e - f\|_2^2 = 1 + 1 - 2e^T f = 2(1 - \cos\theta)$$

Minimizing distance = maximizing cosine = minimizing angle between column spaces

Lagrangean:

$$\mathcal{L}(e,f,v,w,a,b,\alpha,\beta) = 1 - e^T f + a^T(e - Xv) + b^T(f - Yw)$$
$$-\alpha(1 - e^T e) - \beta(1 - f^T f)$$

resulting in

$$\begin{aligned} -f + a &= -e\alpha & \qquad e &= Xv \\ -e + b &= -f\beta & \qquad f &= Yw \\ X^T a &= 0 & \qquad e^T e &= 1 \\ Y^T b &= 0 & \qquad f^T f &= 1 \end{aligned}$$

Eliminating a, b, e, f gives

$$\begin{aligned} X^T Y w &= X^T X v \alpha \\ Y^T X v &= Y^T Y w \beta \end{aligned}$$

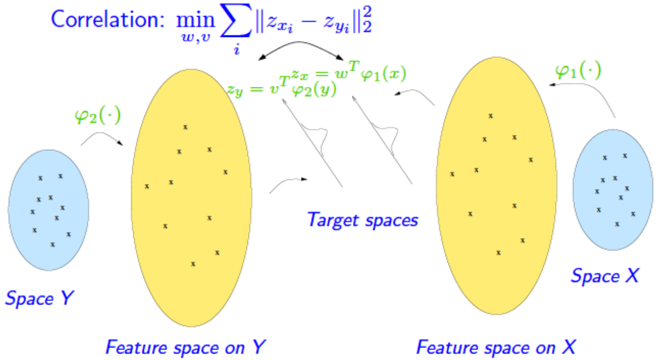
Hence: $\alpha = \beta = \lambda(\text{say})(= \cos\theta)$.

Principal angles and directions follow from Generalized EVP

$$\begin{pmatrix} 0 & X^TY \\ Y^TX & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix} = \begin{pmatrix} X^TX & 0 \\ 0 & Y^TY \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix} \lambda$$

$$v^TX^TXv = w^TY^TYw = 1$$

Numerically correct way: use 3 SVD's (see Golub/VanLoan)



Applications of kernel CCA [Suykens et al., 2002, Bach & Jordan, 2002] e.g. in:
- bioinformatics (correlation gene network - gene expression profiles) [Vert et al., 2003]
- information retrieval, fMRI [Shawe-Taylor et al., 2004]
- state estimation of dynamical systems, subspace algorithms [Goethals et al., 2005]

Kernel CCA

- **Kernel CCA**: primal formulation [Suykens et al., 2002] (related work [Bach & Jordan, 2002])

$$\min_{w,v,b,d,e,r} \; w^T w + v^T v + \nu \sum_i (e_i - r_i)^2 \text{ s.t. } \left\{ \begin{array}{lcl} e_i & = & w^T \varphi_1(x_i) + b, \forall i \\ r_i & = & v^T \varphi_2(z_i) + d, \forall i \end{array} \right.$$

  - Data $\{x_i\}$: **past** *of time-series*
  - Data $\{z_i\}$: **future** *of time-series*
  - **State vector sequence from kernel CCA**
  - System order estimate from kernel CCA

- Dual problem: generalized eigenvalue problem

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])

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

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

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]

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

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

Outline

- Regression

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

- Classification

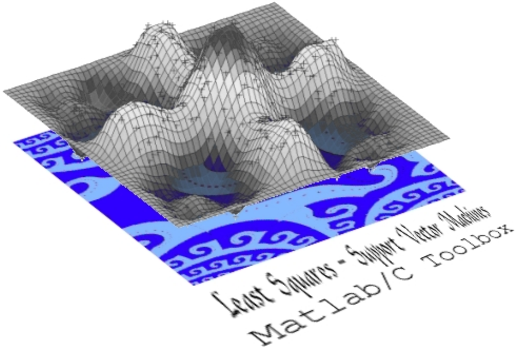
$$\min_{w,b,e} w^T w + \gamma \sum_i e_i^2 \;\; \text{s.t.} \;\; y_i(w^T \varphi(x_i) + b) = 1 - e_i, \;\; \forall i$$

- Kernel pca $(V = I)$, Kernel spectral clustering $(V = D^{-1})$

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

- Kernel canonical correlation analysis/partial least squares

$$\min_{w,v,b,d,e,r} w^T w + v^T v + \nu \sum_i (e_i - r_i)^2 \; \text{s.t.} \left\{ \begin{array}{lcl} e_i & = & w^T \varphi_1(x_i) + b \\ r_i & = & v^T \varphi_2(y_i) + d \end{array} \right.$$

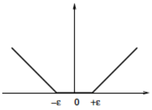


http://www.esat.kuleuven.be/sista/lssvmlab/

Sparsity

- through loss function: model $\hat{y} = \sum_i \alpha_i K(x, x_i) + b$

$$\min \; w^T w + \gamma \sum_i L(e_i)$$



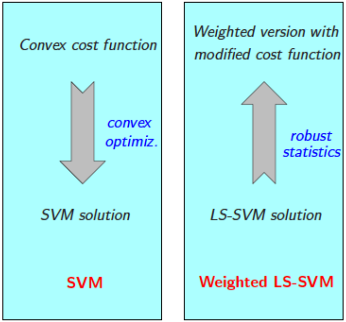
$\Rightarrow$ sparse α

- through regularization: model $\hat{y} = w^T x + b$

$$\min \sum_j |w_j| + \gamma \sum_i e_i^2$$

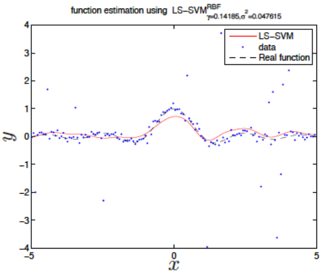
$\Rightarrow$ sparse w

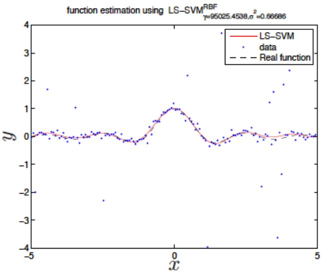
Robustness



- Weighted LS-SVM: $\min_{w,b,e} \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} v_i e_i^2$ s.t. $y_i = w^T \varphi(x_i) + b + e_i, \forall i$
  with v_i determined from $\{e_i\}_{i=1}^N$ of unweighted LS-SVM [Suykens et al., 2002].
  Robustness and stability [Debruyne et al., JMLR 2008, 2010].
- SVM solution by applying iteratively weighted LS [Perez-Cruz et al., 2005]

Example: robust regression using weighted LS-SVM





using LS-SVMlab v1.8 http://www.esat.kuleuven.be/sista/lssvmlab/

Fixed-size method

- Find finite dimensional approximation to feature map $\tilde{\varphi}(\cdot) : \mathbb{R}^p \to \mathbb{R}^M$ based on the eigenvalue decomposition of the kernel matrix (on a **subset** of size $M \ll N$).
- Based on [Williams & Seeger, 2001]:
relates KPCA to a Nyström approximation of the integral equation

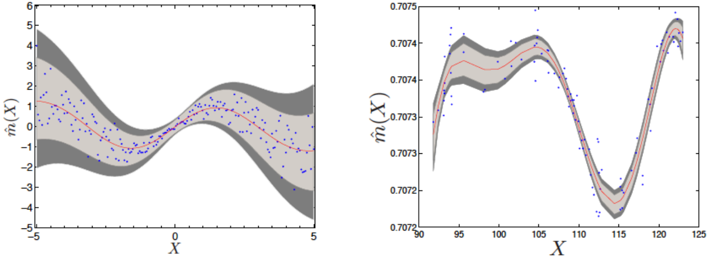
$$\int K(z,x)\phi_i(x)dP_X = \lambda_i \phi_i(z)$$

- Fixed-size method [Suykens et al., 2002; De Brabanter et al., 2009]:
  - selects subset such that it represents the data distribution P_X
  - optimizes quadratic Renyi entropy citerion (instead of random subset)
  - estimate in **primal** by ridge regression (**sparse** representation):

$$\min_{\tilde{w},b} \frac{1}{2}\tilde{w}^T\tilde{w} + \gamma\frac{1}{2}\sum_{i=1}^{N}(y_i - \tilde{w}^T\tilde{\varphi}(x_i) - b)^2$$

KATHOLIEKE UNIVERSITEIT LEUVEN ESAT SCD

Prediction intervals



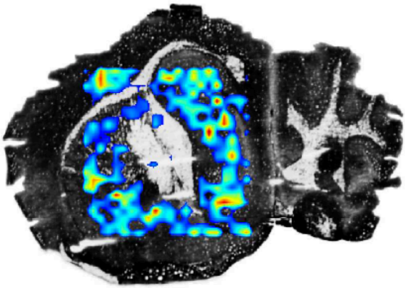
Pointwise and simultaneous 95% prediction intervals for LS-SVM model

[De Brabanter K. et al., IEEE-TNN, 2011], from LS-SVMlab v1.8

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

Tensor completion

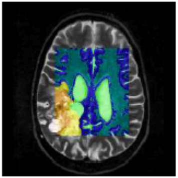


Mass spectral imaging: sagittal section mouse brain [data: E. Waelkens, R. Van de Plas]
Tensor completion using nuclear norm regularization [Signoretto et al., IEEE-SPL, 2011]

Outline

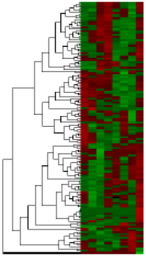
High-quality predictive models are crucial

biomedical



energy

process industry

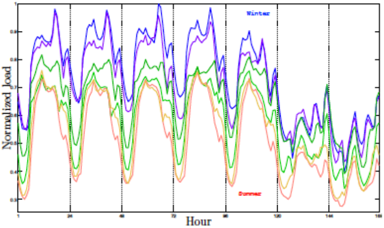


bio-informatics

brain-computer interfaces

traffic networks

Electric Load Forecasting

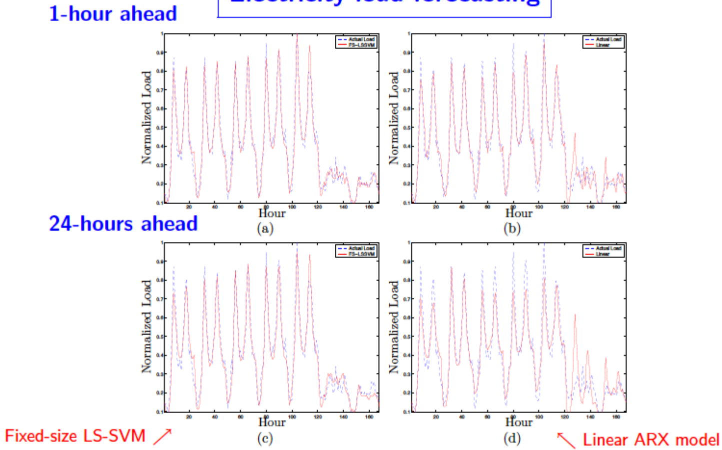


Short-term load forecasting, important for power generation decisions
Hourly load from substations in Belgian grid (ELIA transmission operator)
Seasonal/weekly/intra-daily patterns [Espinoza et al., IEEE CSM 2007]

NARX and AR-NARX model structures: 98 explanatory variables:

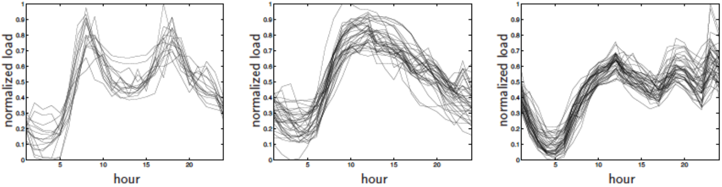
- *lagged load values previous two days (48)*
- *effect of temperature on cooling and heating requirements (3)*
- *calendar information: month, day, hour indications (43)*

Electricity load forecasting



[Espinoza, Suykens, Belmans, De Moor, IEEE CSM 2007]

Power grid: kernel spectral clustering of time-series

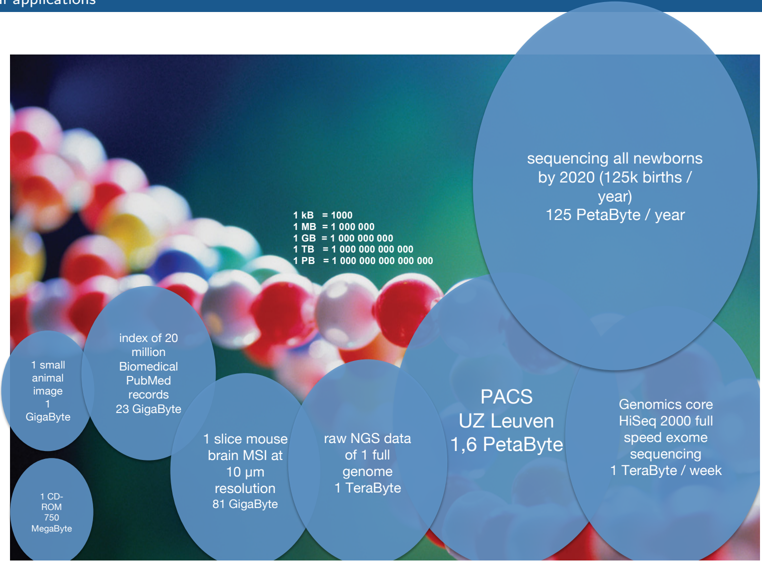


Electricity load: 245 substations in Belgian grid (1/2 train, 1/2 validation)
$x_i \in \mathbb{R}^{43.824}$: spectral clustering on **high dimensional data** (5 years)

3 of 7 detected clusters:
- *1: Residential profile: morning and evening peaks*
- *2: Business profile: peaked around noon*
- *3: Industrial profile: increasing morning, oscillating afternoon and evening*

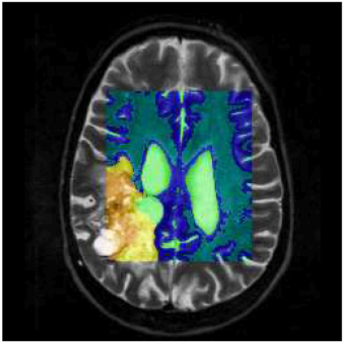
[Alzate, Espinoza, De Moor, Suykens, 2009]

Medical applications



KATHOLIEKE UNIVERSITEIT LEUVEN ESAT SCD

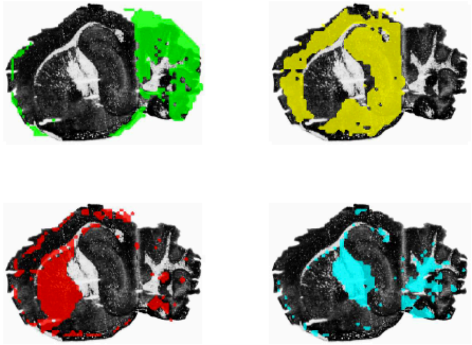
Magnetic resonance spectroscopic imaging



Multiclass LS-SVM classifier:
white matter
gray matter
CSF
grade II glioma
grade III glioma

[Luts J., Ojeda F., Van de Plas R., De Moor B., Van Huffel S., Suykens J.A.K., ACA 2010]

Proteomics



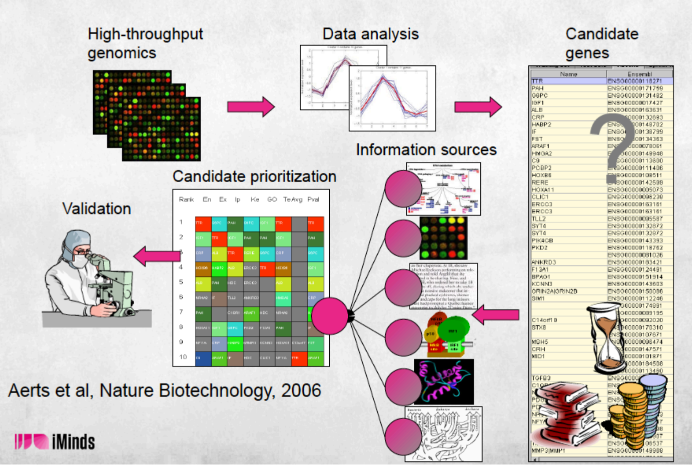
Mass spectral imaging (MSI): section of mouse brain

SVM prediction on 1734 mass spectra (6490 variables/spectrum, 279 pixels, 4 classes)

cerebellar cortex - Ammon's horn section of hippocampus - cauda-putamen - lateral ventricle area

[Luts et al., ACA 2010]

Ranking from data fusion



Outline

- LS-SVM = unifying framework for (un-)supervised ML tasks: regression, (predictive) modeling, clustering and classification, data dimensionality reduction, correlation analysis (spatial-temporal modeling), feature selection, (early - intermediate - late) data fusion, ranking, outlier detection
- Form a core ingredient of decision support systems with 'human decision maker in-the-loop': Policies in climate, energy, pollution; Clinical decision support: digital health; Industrial decision support: yield, monitoring, emission control; Zillions of application areas;
- Tsunami of Big Data (high dimensional input spaces, high complexity and interrelations, ...) are generated by Internet-of-Things multi-sensor networks, clinical monitoring equipment, etc...
- Via the Kernel Trick: *It's all linear algebra !*

STADIUS - SPIN-OFFS "Going beyond research"

www. esat.kuleuven.be/stadius/spinoffs.php

www.tmleuven.be

Data mining
industry solutions

www.dsquare.be

Data handling & mining for
clinical genetics

www.cartagenia.com

Automated PCR analysis

www.ugentec.com

www.baesystems.com

www.ipcos.be