NIPS
NIPS
NIPS
NIPS
NIPS

FEATURE

EXTRACTION

SELECTION

Feature Selection with the Potential Support Vector Machine

Sepp Hochreiter

Technische Universität Berlin

Fakultät für Elektrotechnik und Informatik

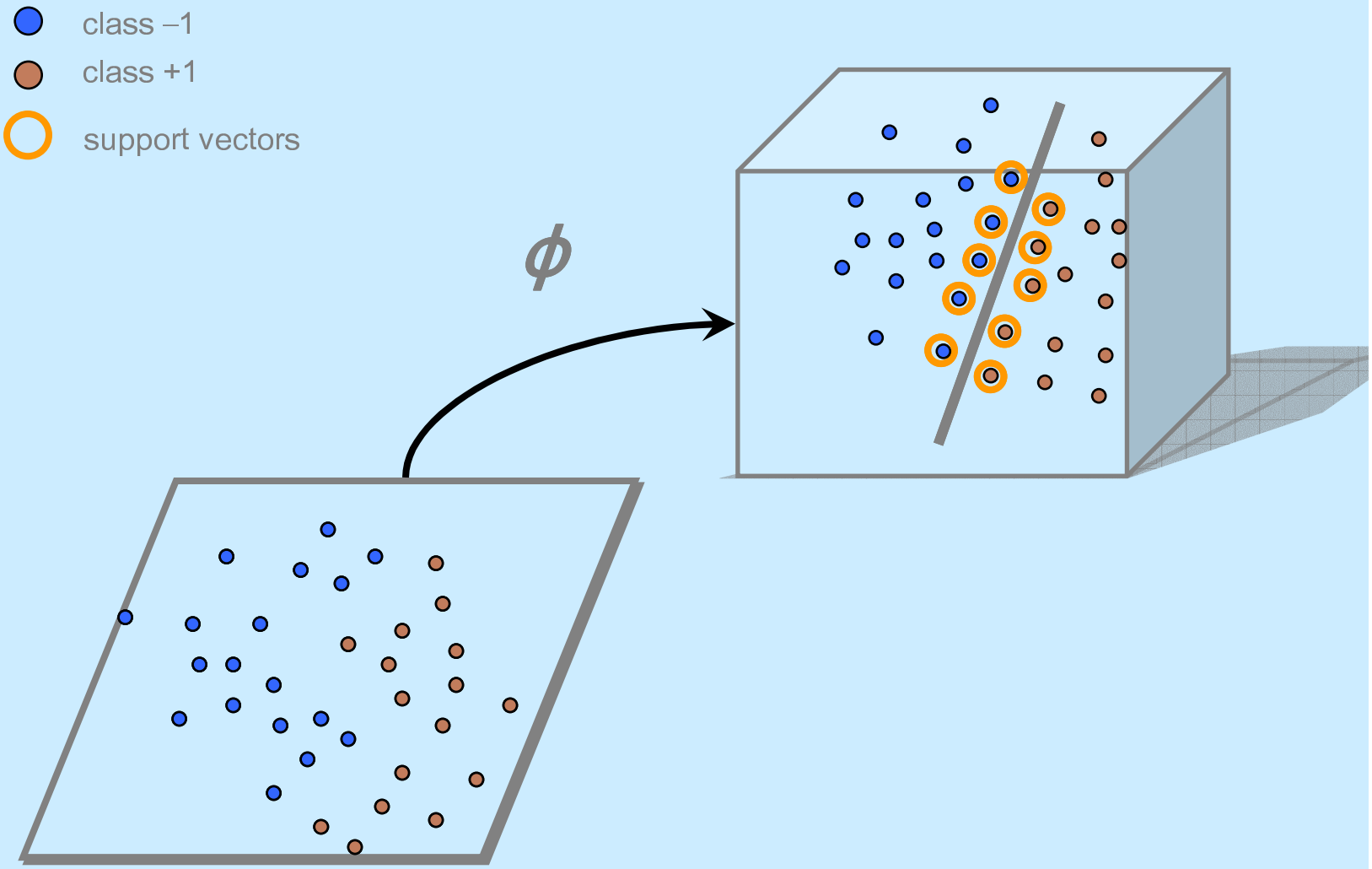
NIPS
NIPS
NIPS
NIPS
NIPS
NIPS

Non-linear Support Vector Machine

FEATURE

EXTRACTION

SELECTION



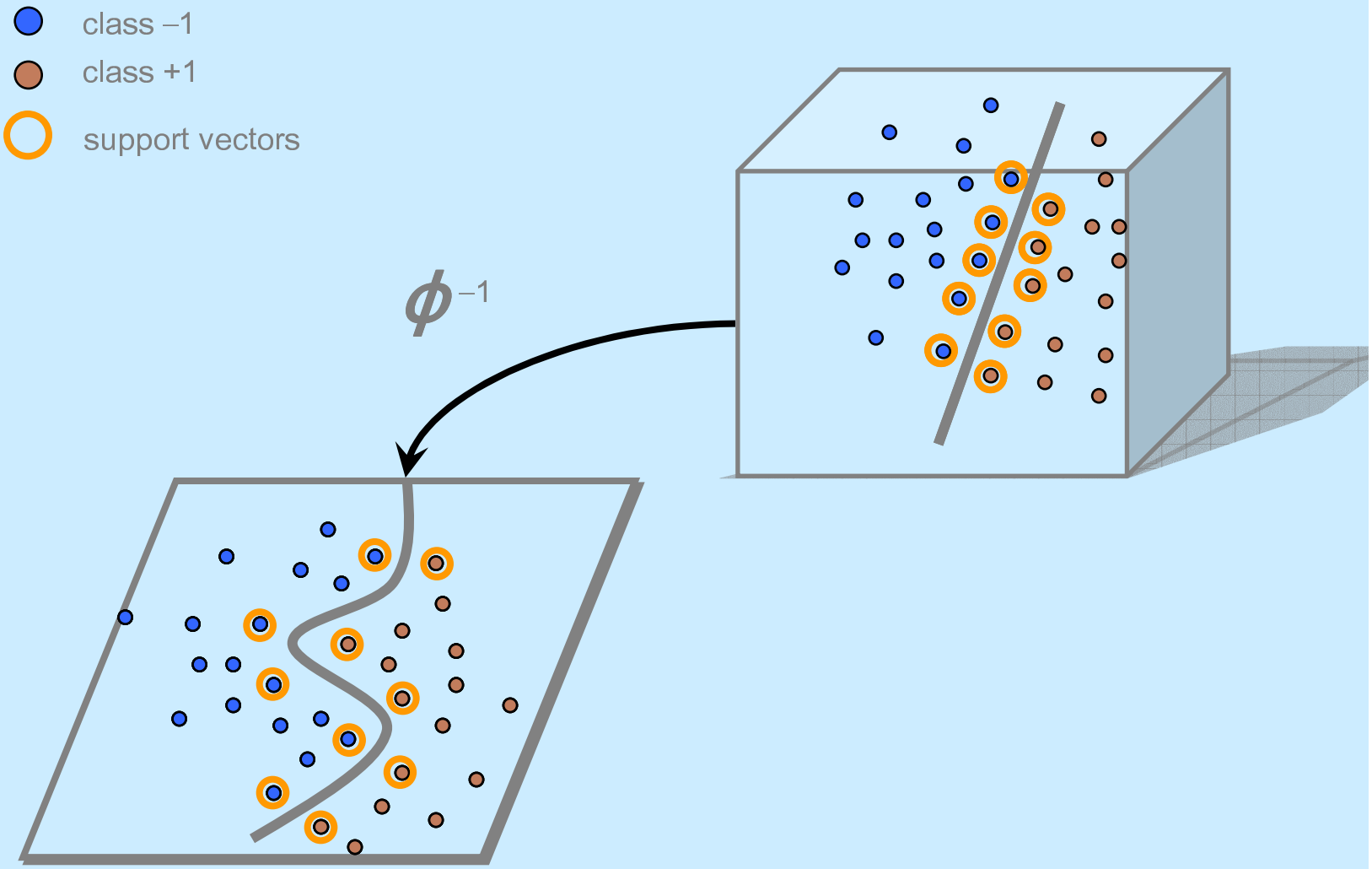
NIPS
NIPS
NIPS
NIPS
NIPS

Non-linear Support Vector Machine

FEATURE

EXTRACTION

SELECTION



NIPS
NIPS
NIPS
NIPS
NIPS

Non-linear Support Vector Machine

FEATURE

EXTRACTION

SELECTION

Kernel Trick

„Kernel Trick“ replaces the dot product with kernel k:

$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ (ϕ may be unknown)

Kernel matrix $\mathbf{K}$, $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, is sufficient for model selection.

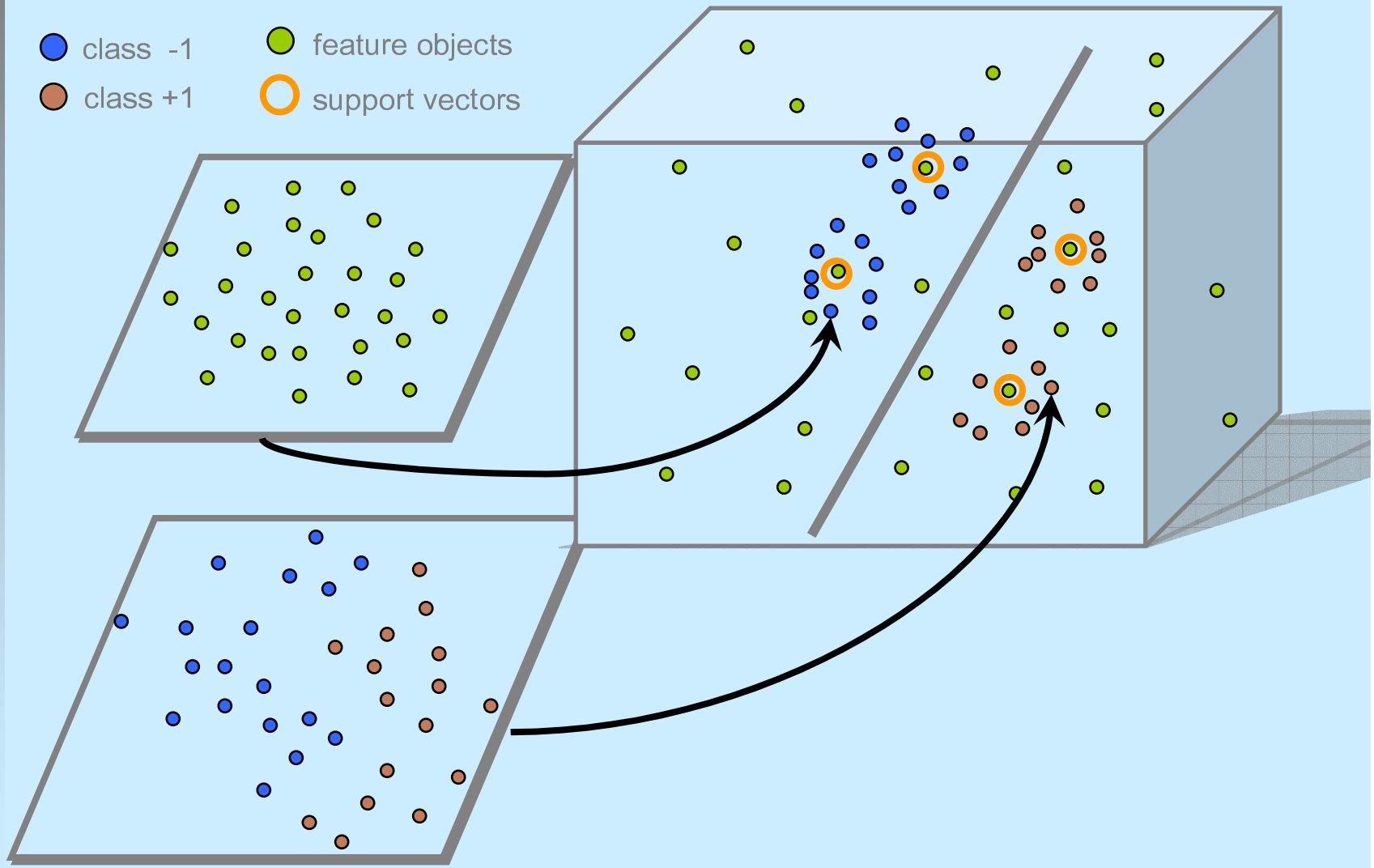
NIPS
NIPS
NIPS
NIPS
NIPS

Basic Idea for Feature Selection

FEATURE

EXTRACTION

SELECTION



NIPS
NIPS
NIPS
NIPS
NIPS

FEATURE

EXTRACTION

SELECTION

Basic Idea for Feature Selection

Consequences

➥Two sets of objects: objects to classify and complex feature objects

➥Both object sets are mapped into the same space (2 mappings)

➥Expansion of the normal vector with respect to the feature objects

→feature weighting (SVM weights the objects to classify)
→ feature extraction

Problem

➥Features are associated with vectors: which vectors?

Solution

➥Data matrix (objects x features) is dot product matrix between complex feature vectors and object vectors

NIPS
NIPS
NIPS
NIPS
NIPS

FEATURE

EXTRACTION

SELECTION

Basic Idea for Feature Selection

Problems with data matrix as dot product matrix

➥Kernel matrix $\mathbf{K}$ may be not positive definite and not squared (SVM optimization is not possible)

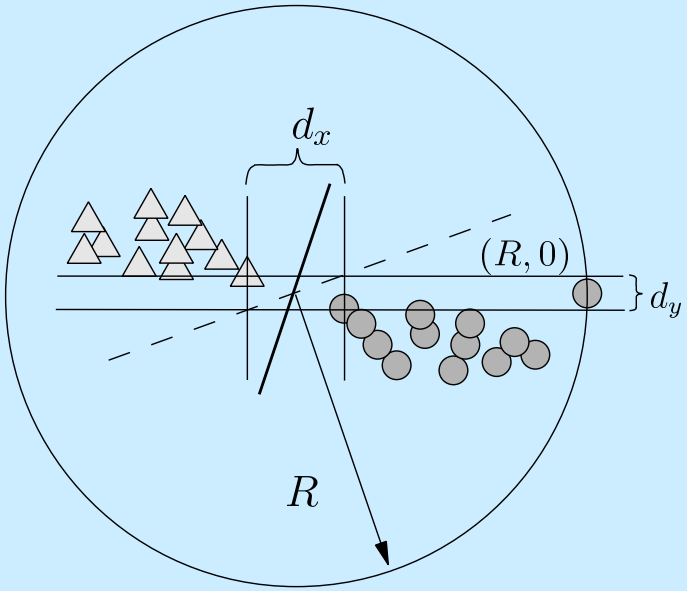
➥Dot products between objects to classify are unknown

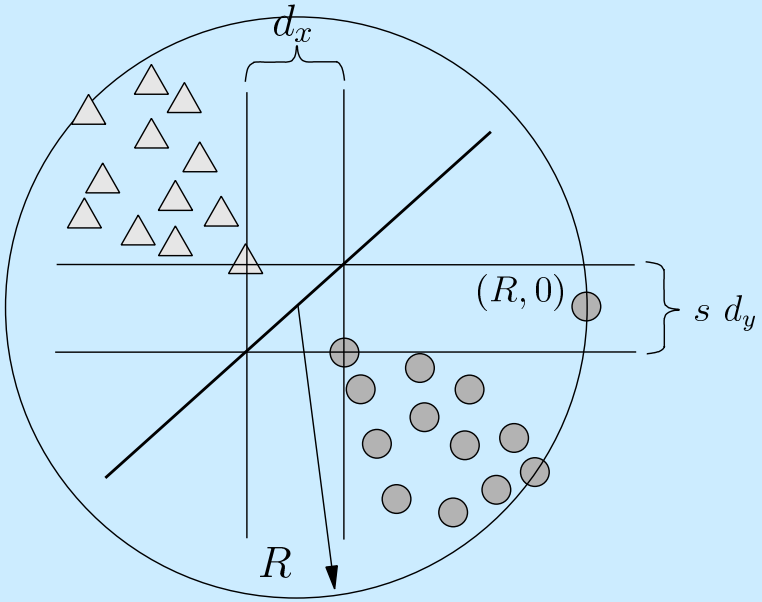
➥SVM technique cannot be applied because $\|\mathbf{w}\|$ cannot be computed

New objective necessary

New constraints necessary

Scale Invariant Objective





SVM solution and error bounds depend on scaling

NIPS
NIPS
NIPS
NIPS
NIPS

FEATURE

EXTRACTION

SELECTION

Scale Invariant Objective

d_x d_x

d_y d_y

$\tilde{R}$

R

Scale invariant objective derived from covering number error bounds:

new $\left\|\mathbf{X}^T\mathbf{w}\right\|_2^2$

SVM $\left\|\mathbf{w}\right\|_2^2$

$\left\|\mathbf{X}^T\mathbf{w}\right\|_2^2 = \mathbf{w}^T\mathbf{X}\,\mathbf{X}^T\mathbf{w}$ ($\mathbf{X}$ is matrix of vectors $\mathbf{x}^i$):

New Constraints

Residual error w.r.t. classification function:

$$r_\alpha \;=\; (w^T \cdot x^\alpha) + b - y^\alpha$$

Minimize the quadratic loss function:

$$\mathrm{R_{emp}} \;=\; \tfrac{1}{2p} \textstyle\sum_{\alpha=1}^{p} r_\alpha^2 \;\stackrel{!}{=}\; \min$$

$$\mathrm{p}\;\nabla_w R_{\mathrm{emp}} \;=\; X(X^T w + b1 - y) \;=\; 0$$

for a linear classifier

NIPS
NIPS
NIPS
NIPS
NIPS

FEATURE

EXTRACTION

SELECTION

New Constraints

Derivative of R_{emp} with respect to **w** along direction $\mathbf{z_j}$ should be zero:

$$\mathrm{dR_{emp}}(w + t z_j)/dt \;=\; z_j^T \nabla_w R_{\mathrm{emp}} \;=\; z_j^T \;\sum_{\alpha=1}^{p} r_\alpha \; x^\alpha \;=\; 0$$

w takes on R_{emp}'s minimum along $\mathbf{z_j}$

$\mathbf{z_j}$ are the complex feature vectors

All directional constraints in matrix form

$$K^T(X^T w + b1 - y) \;=\; 0$$

where

$$K \;=\; X^T \; Z$$

Number of constraints is now the number of complex features

New Constraints

Measurement noise: constraints may lead to overfitting

→relax the constraints through correlation threshold ε
(correlation between $\mathbf{K}_{ij}$ and $\mathbf{r}_j$)

$$K^T(X^T w + b1 - y) \; - \; \epsilon \; \leq \; 0$$
$$K^T(X^T w + b1 - y) \; + \; \epsilon \; \geq \; 0$$

Normalization of **K** to equal feature variance is necessary to use one ε for all constraints, that is for all complex features

Increase of the residual error after the elimination of the j-th feature is bounded by:

$$2\,\epsilon|w_j| \; + \; pw_j^2$$

NIPS
NIPS
NIPS
NIPS
NIPS

Potential Support Vector Machine

FEATURE

EXTRACTION

SELECTION

Algorithm

Primal

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\left\|\mathbf{X}^T\mathbf{w}\right\|_2^2$$

$$\text{s.t.} \quad \mathbf{K}^T\left(\mathbf{X}^T\mathbf{w}+b\,\mathbf{1}-\mathbf{y}\right)+\varepsilon\mathbf{1} \geq \mathbf{0}$$

$$\mathbf{K}^T\left(\mathbf{X}^T\mathbf{w}+b\,\mathbf{1}-\mathbf{y}\right)-\varepsilon\mathbf{1} \leq \mathbf{0}$$

- $\mathbf{y}$ is vector of labels y_i
- $\mathbf{Z}$ is matrix of feature objects $\mathbf{z}_j$

$$\boxed{\mathbf{K}=\mathbf{X}^T\mathbf{Z}}$$

Lagrangian

$\mathbf{X}\,\mathbf{X}^T\mathbf{w}=\mathbf{X}\,\mathbf{X}^T\mathbf{Z}\,\boldsymbol{\alpha}$ is assured by $\boxed{\mathbf{w}=\mathbf{Z}\,\boldsymbol{\alpha}}$

$\mathbf{w}$ expanded with respect to features

Potential Support Vector Machine

Algorithm

Dual

$$\min_{\alpha^+,\,\alpha^-} \; \frac{1}{2}\left(\alpha^+ - \alpha^-\right)^T \mathbf{K}^T\mathbf{K}\left(\alpha^+ - \alpha^-\right) - \mathbf{y}^T\mathbf{K}\left(\alpha^+ - \alpha^-\right) + \varepsilon\,\mathbf{1}^T\left(\alpha^+ + \alpha^-\right)$$

$$\text{s.t.} \quad \mathbf{1}^T\mathbf{K}\left(\alpha^+ - \alpha^-\right) = 0\,, \qquad C\,\mathbf{1} \geq \alpha^+,\,\alpha^- \geq 0$$

$\boxed{\mathbf{w} = \mathbf{Z}\,\alpha}$, where $\alpha = \alpha^+ - \alpha^-$.

$\mathbf{K}^T\mathbf{K}$ is (features x features) and optimization would be computational expensive: Sequential Minimal Optimization (SMO)

Potential Support Vector Machine

Characteristic

- Works with data matrix
- Feature selection: Identification of relevant features

Applications: Prediction of a treatment outcome based on the gene expression profile obtained from the micro array technique

- Brain tumor
- Breast cancer

NIPS
NIPS
NIPS
NIPS
NIPS

FEATURE

EXTRACTION

SELECTION

Brain Tumor

Task

Brain tumor (medulloblastoma) patients respond differently to the chemotherapy and radiation

➥Negative prognoses: alternative therapy or more intensive control

➥Positive prognoses: toxicity of the therapy can be reduced

60 patients and 7129 genes

S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova and P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. MOlson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, T. R. Golub
Prediction of central nervous system embryonal tumour outcome baed on gene expression
Nature 415(687):436-442, 2002

NIPS
NIPS
NIPS
NIPS
NIPS

Brain Tumor

FEATURE

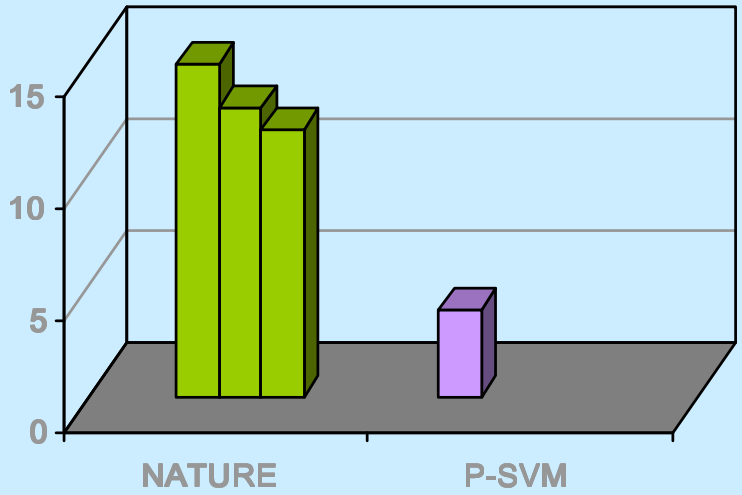
EXTRACTION

SELECTION

Classification results

Standard			New Method (P-SVM)		
Method	F	Error	Method	Features	Error
TrkC (ein Gen)	1	20	SVM	40 / 45 / 50	5 / 4 / 5
SVM		15	SVM	40 / 45 / 50	5 / 5 / 5
TrkC & SVM		14	P-SVM	40 / 45 / 50	4 / 4 / 5
KNN	8	13			
KNN & SVM		12			

Standard feature selection with „signal-to-noise“- and „*t*“-statistic



Breast Cancer

Task

Breast cancer: the treatment is for 70-80 % of the patients not necessary to avoid metastasis

➥ Prediction of metastasis leads to choice of patients for therapy

➥Alternative treatment and toxicity reduction

78 patients und 25000 genes

L. J. van't Veer, H, Dai, M. J. van de Vijver, Y. D. He, A. A. MHart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, S. H. Friend
Gene expression profiling predicts clinical outcome of breast cancer
Nature 415: 530-536, 2002

NIPS
NIPS
NIPS
NIPS
NIPS

Breast Cancer

FEATURE

EXTRACTION

SELECTION

Classification results

Standard Feature Selection					New Method (P-SVM)				
Method	F	Error	ROC	Test	Method	F	Error	ROC	Test
weighted voting	70	20	0.77	2	SVM	30	12	0.88	2

Standard feature selection with „signal-to-noise“-statistic

NIPS
NIPS
NIPS
NIPS
NIPS

FEATURE

EXTRACTION

SELECTION

Feature selection history