Discriminant Analysis for Positive Definite and Indefinite Kernels

COMPSTAT 2010
Paris, 22.-27.8.2010

Bernard Haasdonk
University of Stuttgart, Germany
Institute of Applied Analysis and Numerical Simulation
haasdonk@mathematik.uni-stuttgart.de

Elzbieta Pekalska
University of Manchester
School of Computer Science
pekalska@cs.man.ac.uk

SimTech Cluster of Excellence

Universität Stuttgart

MANCHESTER 1824 The University of Manchester

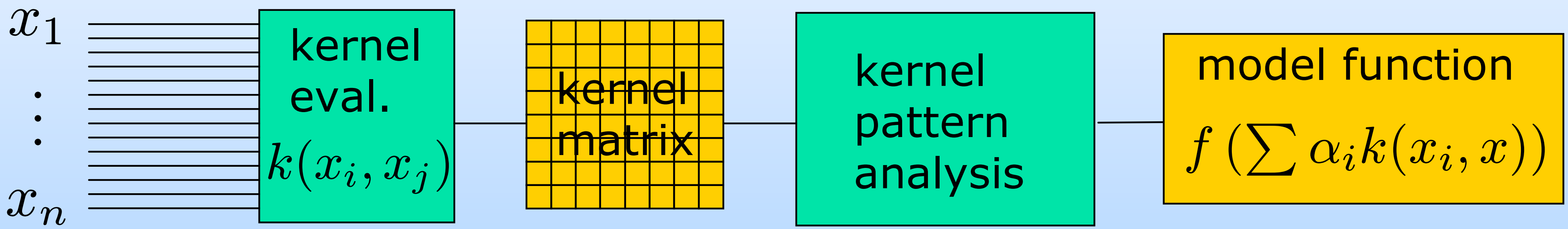
Overview

- Kernel Methods and Indefinite Spaces
  - Kernel Methods
  - Pseudo-Euclidean Spaces
  - Indefinite Support Vector Machine
- Kernel Discriminant Classification
  - Kernel Quadratic Discriminant Classification
  - Indefinite Kernel Fisher Discriminant
- Kernel Discriminant Feature Extraction
  - Indefinite Kernel Mahalanobis distance
  - Indefinite KFDA
- Summary and Conclusion

Kernel Methods and Indefinite Spaces

Kernel Methods [SS02,SC04]

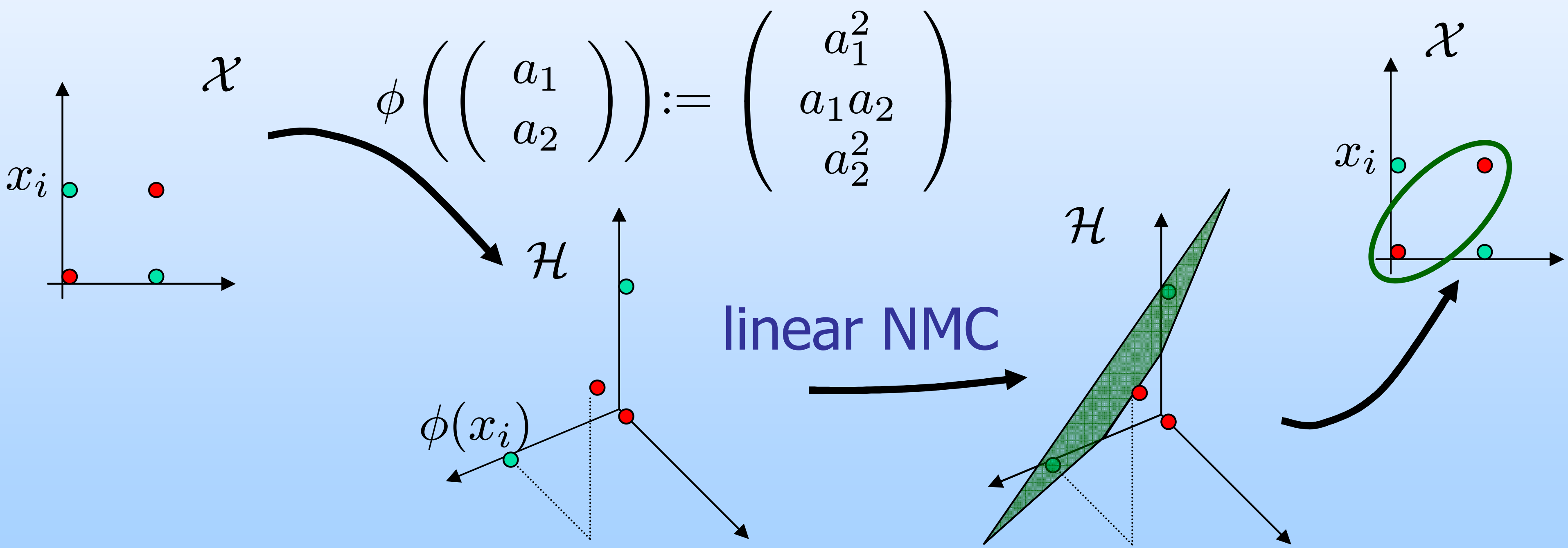
- Typical tasks:
  - Classification, Regression, Clustering, Novelty Detection,…
- Multitude of kernel methods: SVM, SVR, KPCA,…
- Analysis chain:



- Multitude of kernels for various datatypes
  - Vectorial, sequences, graphs, finite state machines…
- Kernel matrix is information „bottleneck“
  => importance of kernel choice!

Geometrical Interpretation

- Choose a mapping $\phi : \mathcal{X} \to \mathcal{H}$ into a Hilbert space $\mathcal{H}$
- Linear method in $\mathcal{H}$ yields a nonlinear method in $\mathcal{X}$



- Kernel function for inner-products $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$
- Mapping no longer required explicitly

Distance Substitution Kernels [HBB04]

- Distance $d(\cdot,\cdot)$ symmetric, nonnegative, zero-diagonal
- Examples of DS-kernels:

$$k_d^{\mathrm{lin}}(x,x') := \langle x,x'\rangle_d^O \qquad k_d^{\mathrm{nd}}(x,x') := -d(x,x')^\beta, \beta \in [0,2]$$

$$k_d^{\mathrm{pol}}(x,x') := \left(1+\gamma\langle x,x'\rangle_d^O\right)^p \qquad k_d^{\mathrm{rbf}}(x,x') := \mathrm{e}^{-\gamma d(x,x')^2}, p \in \mathbb{N}, \gamma \geq 0$$

where $O \in \mathcal{X}$ is an arbitrary origin and

$$\langle x,x'\rangle_d^O := -\frac{1}{2}\left(d(x,x')^2 - d(x,O)^2 - d(x',O)^2\right)$$

- Expectation: Similar behaviour as standard kernels
- Generality: Arbitrary structured Objects + Distances!
- (c)pd-ness equivalent to d being a Hilbertian metric

Distance Substitution Kernels

- Many DS-Kernels are positive definite
- Examples of Hilbertian Metrics:
  - Hellinger Distance

$$(H(p,p'))^2 := \int (\sqrt{p} - \sqrt{p'})^2 dx$$

  - Chi-Square

$$\chi^2(\mathbf{x},\mathbf{y}) := \frac{1}{2}\sum_i \frac{(x_i - y_i)^2}{x_i + y_i}$$

  - Powers of p-norms

$$\|\mathbf{x} - \mathbf{x}'\|_p^q \qquad p \in [0,2], q \in [0, p/2]$$

  - Variation of Kulback Leibler:

$$d^2_{1|1}(p,p') := \frac{1}{2}\int_{\mathcal{X}} p(x)\log\left(\frac{2p(x)}{p(x)+p'(x)}\right) + p'(x)\log\left(\frac{2p'(x)}{p(x)+p'(x)}\right) d\mu(x)$$

Indefinite Kernels

- Sources of Indefiniteness
  - Distance-based kernels: non-Hilbertian, non-metric
  - Prior knowledge in kernel construction
  - Invariant kernels
  - Robust or approximate (dis)similarities
  - Kernel combination
- Indefinite Kernel Methods
  - Nearest Mean Classifier [PD05]
  - Regression [OMCS04]
  - Indefinite Support Vector Machine [H05b]
  - Indefinite Fisher Discriminant [HP08b]
  - Indefinite Kernel Quadratic Discriminant [PH09]
  - Kernel Mahalanobis Distances [HP08,HP10]

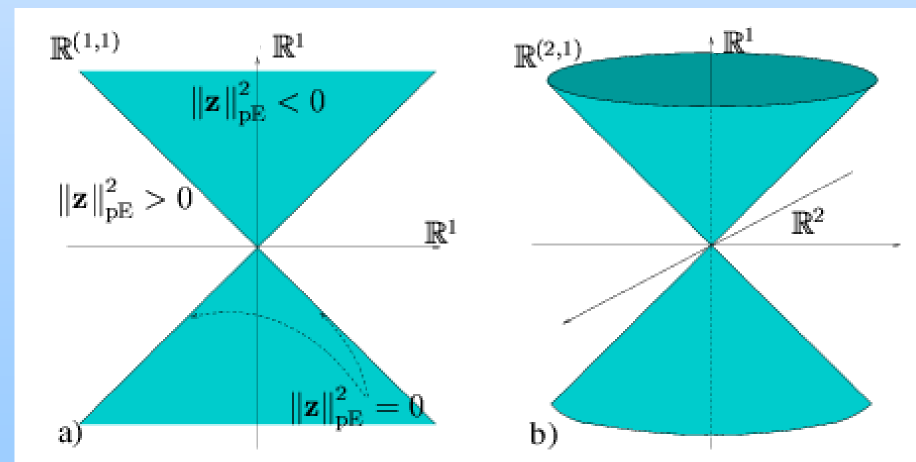
Pseudo-Euclidean Spaces [G85,PPD01]

- Real finite dimensional vector spaces
  $\mathbb{R}^{(p,q)} := \mathbb{R}^p \oplus \mathbb{R}^q$ of signature (p,q)
- symmetric (indefinite) inner-product

$$\langle \mathbf{z}, \mathbf{z}' \rangle_{\mathrm{pE}} := \mathbf{z}_p^T \mathbf{z}'_p - \mathbf{z}_q^T \mathbf{z}'_q = \mathbf{z}^T \mathbf{J} \mathbf{z}' \qquad \mathbf{J} := \mathrm{diag}(\mathbf{1}_p, -\mathbf{1}_q)$$

- squared norm
  $\|\mathbf{z}\|_{\mathrm{pE}}^2 := \langle \mathbf{z}, \mathbf{z} \rangle_{\mathrm{pE}} = \mathbf{z}^T \mathbf{J} \mathbf{z}$ can be negative:
- squared distance
  $\|\mathbf{z} - \mathbf{z}'\|_{\mathrm{pE}}^2 = \langle \mathbf{z} - \mathbf{z}', \mathbf{z} - \mathbf{z}' \rangle_{\mathrm{pE}}$
- orthogonality
  $\langle \mathbf{z}, \mathbf{z}' \rangle_{\mathrm{pE}} = \mathbf{z}^T \mathbf{J} \mathbf{z}' = 0$
- hyperplanes

$$H : \langle \mathbf{z}, \mathbf{w} \rangle_{\mathrm{pE}} + b = 0$$



pE Feature Space Embedding

- Data dependent pE-embedding:

Given data

$\{x_i\}_{i=1}^n \subset \mathcal{X}$

+ sym. kernel

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

⇒

Existence of pE space $\mathbb{R}^{(p,q)}$

+ embedding $\Phi : \mathcal{X} \to \mathbb{R}^{(p,q)}$

with $k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathrm{pE}}$

- Construction by Eigendecomposition [GHBO99,PPD01]:

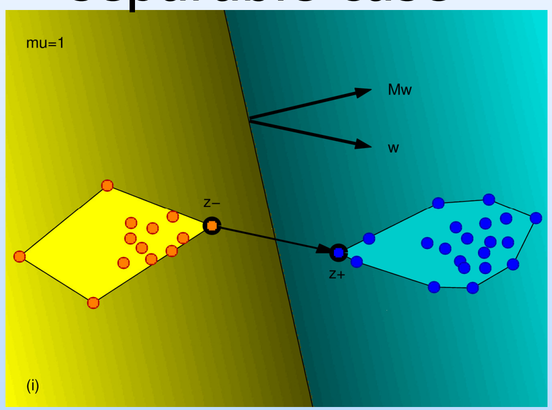
$\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$ $\quad p := \dim(\boldsymbol{\lambda}^+), q := \dim(\boldsymbol{\lambda}^-)$

$\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda}^+, \boldsymbol{\lambda}^-)$ $\quad \Phi(x_i) := \left(\sqrt{|\boldsymbol{\Lambda}|}\mathbf{U}^T\right)_i$

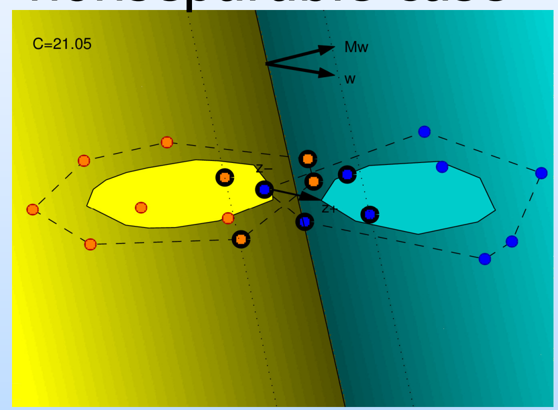
Indefinite SVM [Ha05]

- Geometric interpretation: optimal hyperplane classifier
  - not margin maximization but separation of convex hulls

separable case



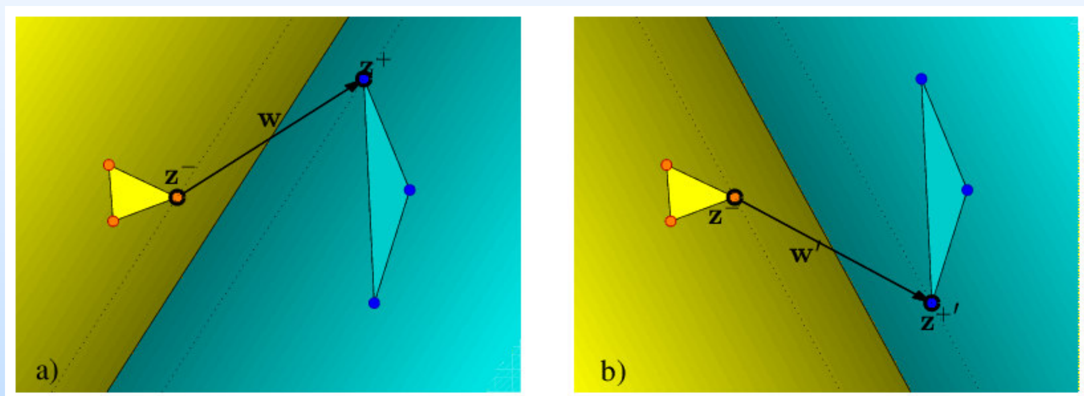
nonseparable case



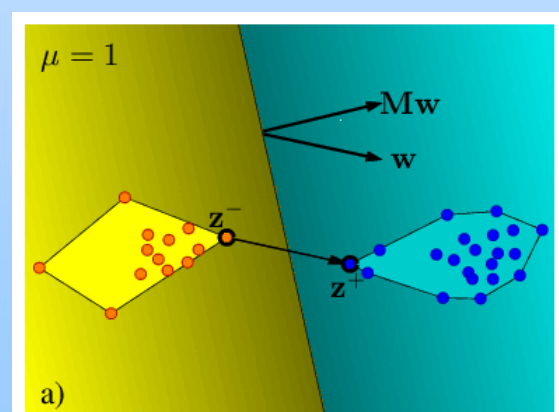
- Sparseness: usual interpretation of support vectors
  - E.g: $\alpha_i = 0 \quad \Rightarrow$ sample x_i is correctly classified
- Numerics: convergence, e.g. libsvm [LL03]
- Uniqueness: possible but generally not
- Suitability criteria: e.g. $\mathbf{w}^T \mathbf{M} \mathbf{w}$, #bSV, DCM

Numerics of Indefinite SVM

- Convergence to stationary point, libsvm [LL03]
- Multiple solutions

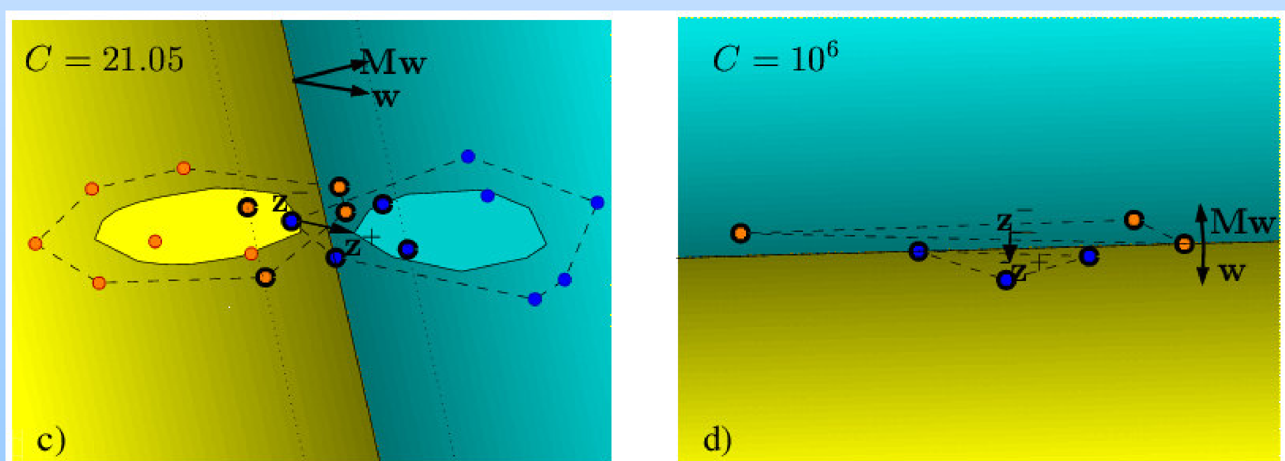


- Uniqueness in extreme indefinite cases



Practical Criteria for Indefinite SVM

- Criterion for suitability: #bSV
  - No (few) bounded $\alpha_i \Rightarrow$ no (few) training errors
- Criterion for unsuitability: $\mathbf{w}^T\mathbf{M}\mathbf{w} \leq 0$
  - after training: $\sum_{i,j} \alpha_i\alpha_j y_i y_j k(x_i, x_j)$
  - before training:
    high negative signature of the pE space
- Criterion for suitability: Distance of Class Means
  - If DCM is positive, sufficiently low C yields solution



$$\mathrm{DCM}^2 = \sum_{i,j} c_i c_j k(x_i, x_j)$$

$$c_i = \begin{cases} 1/n^+ & \text{for } y_i = +1 \\ -1/n^- & \text{for } y_i = -1 \end{cases}$$

Kernel Quadratic Discriminant Classifier

Quadratic Discriminant Analysis [DHS01]

- Multiclass problem $\Omega := \{\omega_1, \ldots, \omega_c\}$, patterns $x \in \mathbb{R}^k$
- Class-conditional normal densities

$$p(x|\omega_j) = \mathcal{N}(x; \{\Sigma^{[j]}, \mu^{[j]}\})$$
$$= ((2\pi)^k \det(\Sigma^{[j]}))^{-1/2} \exp\left(-\tfrac{1}{2}(x-\mu^{[j]})^T (\Sigma^{[j]})^{-1}(x-\mu^{[j]})\right)$$

- MAP decision functions

$$f_j(x) = -\tfrac{1}{2}(x-\mu^{[j]})^T (\Sigma^{[j]})^{-1}(x-\mu^{[j]}) + b_j$$
$$b_j = -\tfrac{1}{2}\ln(\det(\Sigma^{[j]})) + \ln(P(\omega_j))$$

- QD classification by maximal decision functions

x assigned class ω_i if $i = \arg\max_{1 \le j \le c} f_j(x)$

⇨ Goal: Kernelization of Mahalanobis distance + bias

Basic Notation

- Training samples $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \Omega$
- Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and RKHS-embedding $\phi : \mathcal{X} \to \mathcal{H}$

  $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}, \qquad \Phi := (\phi(x_1), \ldots, \phi(x_n))$
- Kernel matrix — kernel vector

  $K := (\langle \phi(x_i), \phi(x_j) \rangle)_{i,j=1}^n =: \Phi^T \Phi \qquad \mathbf{k}_x := (k(x_i, x))_{i=1}^n$
- Mean $\phi_\mu := \frac{1}{n} \Phi \mathbf{1}_n$ and centering $\tilde{\phi}(x) := \phi(x) - \phi_\mu$

  $\tilde{\Phi} := \Phi - \phi_\mu \mathbf{1}_n^T, \qquad \tilde{K} := \tilde{\Phi}^T \tilde{\Phi} \qquad \tilde{\mathbf{k}}_x := \tilde{\Phi}^T \tilde{\phi}(x)$
- Empirical covariance operator $C : \mathcal{H} \to \mathcal{H}$ acting as

  $$Cv := \frac{1}{n} \sum_{i=1}^n (\phi(x_i) - \phi_\mu) \langle \phi(x_i) - \phi_\mu, v \rangle_{\mathcal{H}} = \frac{1}{n} \tilde{\Phi} \tilde{\Phi}^T v$$
- Class-wise quantities: superscript $[j]$

Kernel Quadratic Discriminant KQD-IC

- Assumption: **I**nvertible **C**ovariance operator
- W.l.o.g. finite dimensional $\mathcal{H} = \mathbb{R}^m, m < n$
- SVD of $\tilde{\Phi} \in \mathbb{R}^{m \times n}$ yields kernel/covariance relation

$$\frac{1}{n} C^{-1} \tilde{\Phi} = \tilde{\Phi} \tilde{K}^{-}$$

- Covariance operation is

$$C \tilde{\phi}(x) = \frac{1}{n} \tilde{\Phi} \tilde{\mathbf{k}}_x$$

- Kernelized Mahalanobis distance follows as

$$\tilde{\phi}(x)^T C^{-1} \tilde{\phi}(x) = n \tilde{\mathbf{k}}_x^T (\tilde{K}^{-})^2 \tilde{\mathbf{k}}_x$$

Kernel Quadratic Discriminant KQD-IC

- KQD-**IC** decision function:

$$f_j(x) = -\frac{n_j}{2}(\tilde{\mathbf{k}}_x^{[j]})^T((\tilde{K}^{[j]})^-)^2\tilde{\mathbf{k}}_x^{[j]} + b_j$$

with $(\tilde{K}^{[j]})^-$ pseudo inverse of $\tilde{K}^{[j]}$

$\mathbf{k}_x^{[j]} := (k(x_i^{[j]}, x))_{i=1}^{n_j}$ kernel vector

$\tilde{\mathbf{k}}_x^{[j]} := H^{[j]}(\mathbf{k}_x^{[j]} - \frac{1}{n_j}K^{[j]})\mathbf{1}_{n_j}$ centered kernel vector

$H^{[j]} := I_{n_j} - \frac{1}{n_j}\mathbf{1}_{n_j}\mathbf{1}_{n_j}^T$ centering matrix

- Regularization parameter $\alpha_j > 0$ of pseudo-inverse:

eigenvalues $\lambda_i^{[j]}$ with $|\lambda_i^{[j]}| < \alpha_j$ set to 0

Kernel Quadratic Discriminant KQD-RC

- Ansatz: **R**egularization of **C**ovariance operator
- No restriction on dimensionality of $\mathcal{H}$
- $C_{reg} := C + \sigma^2 I$ gives kernel/covariance relation

$$\frac{1}{n} C_{reg}^{-1} \tilde{\Phi} = \tilde{\Phi} \tilde{K}_{reg}^{-1}$$

  by setting $\tilde{K}_{reg} := \tilde{K} + n\sigma^2 I_n$
- Covariance operation is

$$C_{reg} \tilde{\phi}(x) = \frac{1}{n} \tilde{\Phi} \tilde{\mathbf{k}}_x + \sigma^2 \tilde{\phi}(x)$$

- Kernelized Mahalanobis distance follows as

$$\tilde{\phi}(x)^T C_{reg}^{-1} \tilde{\phi}(x) = \frac{1}{\sigma^2} (\tilde{k}_{xx} - (\tilde{\mathbf{k}}_x)^T (\tilde{K}_{reg})^{-1} \tilde{\mathbf{k}}_x)$$

Kernel Quadratic Discriminant KQD-RC

- KQD-**RC** decision function:

$$f_j(x) = -\frac{1}{2\sigma_j^2}(\tilde{k}_{xx}^{[j]} - (\tilde{\mathbf{k}}_x^{[j]})^T(\tilde{K}_{reg}^{[j]})^{-1}\tilde{\mathbf{k}}_x^{[j]}) + b_j$$

with $\tilde{K}_{reg}^{[j]} := \tilde{K}^{[j]} + n_j\sigma_j^2 I_{n_j}$ regularized kernel matrix

$$\tilde{k}_{xx}^{[j]} := k_{xx} - \frac{2}{n_j}\mathbf{1}_{n_j}^T\mathbf{k}_x^{[j]} + \frac{1}{n_j^2}\mathbf{1}_{n_j}^T K^{[j]}\mathbf{1}_{n_j}$$

$$k_{xx} := k(x,x)$$

- Regularization parameter $\sigma_j > 0$
guarantees regularized kernel matrix to be invertible

Bias Computation

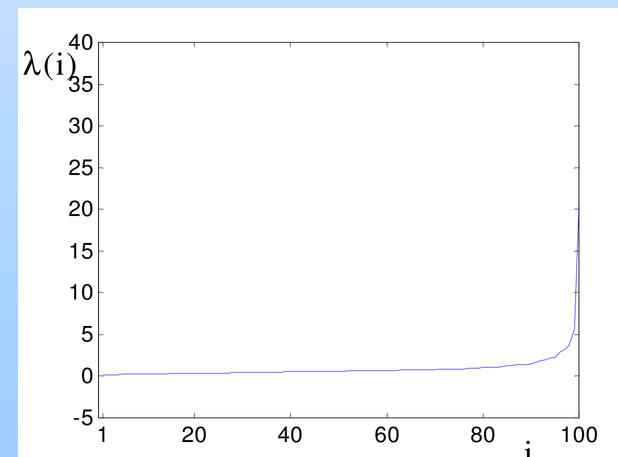
- Kernelized bias:
  - Assumption: regularized covariance

$$b_j = -\frac{1}{2}\ln(\det(C_{reg}^{[j]})) + \ln(P(\omega_j))$$

$$\ln(\det(C_{reg}^{[j]})) = \ln(\prod_{i=1}^{l} \lambda_i^{[j]}) = \sum_{i=1}^{l} \ln(\lambda_i^{[j]})$$

  - Eigenvalues $\lambda_i^{[j]}$ of $C_{reg}^{[j]}$ obtained as those of $\frac{1}{n_j}\tilde{K}_{reg}^{[j]}$
  - Similar for invertible covariance operators
- Problem: numerical instability
  - long eigenvalue tails, many small eigenvalues
  - Estimate of intrinsic dimensionality required

⇨ high variation of bias computation!!



Bias Computation

- Solution
  - QD is the Bayes classifier in case of known statistics
    Hence, bias values minimize prediction error
  - Use training error as surrogate for prediction error
- Bias computation: training error minimization
  - KQD decision invariant to simultaneous shifting of biases
  - For 2-class decision only difference of biases is relevant
  - Estimate optimal pairwise bias differences for all classes

$$\Delta_{ij} \approx b_i - b_j$$

    by training error minimization and greedy search
  - Solve overal least squares problem for biases

$$\min_{\mathbf{b}} \sum_{i=1}^{c-1} \sum_{j=i+1}^{c} (b_i - b_j - \Delta_{ij})^2$$

Experiments

- Real world data
  - Data from [ROM98]: superiority of KFD
  - 2 classes, 2-60 dimensions, 215-7400 samples
  - Gaussian Kernel, 10-fold CV for regularization & kernel

10-fold repetition, test-errors: mean (std)

	Banana	Breast-cancer	Diabetis	Flare-solar	German	Heart	Image
KQD-IC	11.7 (0.5)	35.4 (4.2)	28.0 (2.6)	35.4 (2.5)	30.4 (2.4)	21.6 (5.3)	3.3 (0.7)
KQD-RC	12.1 (0.2)	39.0 (3.7)	30.9 (1.8)	34.2 (1.3)	28.2 (2.4)	19.1 (3.4)	3.3 (0.6)
KFD	11.8 (0.4)	34.4 (4.2)	26.9 (1.8)	33.6 (2.0)	27.1 (2.2)	18.4 (2.8)	2.8 (0.7)
KNN	12.4 (0.2)	40.7 (5.9)	33.9 (2.7)	34.1 (2.2)	36.0 (2.7)	19.3 (3.7)	3.6 (0.4)
KPCA-QD	12.0 (0.5)	36.3 (5.4)	29.1 (2.1)	32.5 (2.4)	28.5 (2.4)	19.7 (2.9)	5.1 (1.0)
	Ringnorm	Splice	Thyroid	Titanic	Twonorm	Waveform	
KQD-IC	2.9 (0.7)	16.1 (1.0)	5.8 (3.6)	33.7 (3.3)	3.4 (0.2)	14.3 (1.3)	
KQD-RC	1.6 (0.2)	11.8 (1.1)	7.5 (3.4)	32.3 (2.4)	2.6 (0.3)	13.5 (1.7)	
KFD	1.8 (0.2)	10.6 (0.7)	6.8 (3.8)	30.7 (1.9)	2.6 (0.3)	10.1 (0.6)	
KNN	42.7 (10.2)	22.8 (1.1)	10.3 (10.9)	33.8 (4.4)	3.8 (0.3)	12.9 (1.3)	
KPCA-QD	1.8 (0.2)	15.5 (0.8)	8.4 (5.3)	30.3 (1.2)	2.6 (0.4)	12.1 (1.3)	

- No clear favorite among KQD-IC/RC
- KQD outperforming KNN, almost as good as KFD,
- comparable to KPCA-QD

Indefinite Kernel Fisher Discriminant

Pseudo Euclidean Fisher Discriminant

- Class means $\boldsymbol{\mu}_\pm := \frac{1}{n_\pm} \sum_{i \in I_\pm} \phi(x_i)$
- Between-class scatter projection

$$\Sigma^B_{\mathrm{pE}} \mathbf{w} = (\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-) \left\langle \boldsymbol{\mu}_+ - \boldsymbol{\mu}_-, \mathbf{w} \right\rangle_{\mathrm{pE}}$$

- Within-class scatter projection

$$\Sigma^W_{\mathrm{pE}} \mathbf{w} = \Sigma^W_{\mathrm{pE},+} \mathbf{w} + \Sigma^W_{\mathrm{pE},-} \mathbf{w}$$

$$\Sigma^W_{\mathrm{pE},\pm} \mathbf{w} = \sum_{i \in I_\pm} (\phi(x_i) - \boldsymbol{\mu}_\pm) \left\langle \phi(x_i) - \boldsymbol{\mu}_\pm, \mathbf{w} \right\rangle_{\mathrm{pE}}$$

- Maximize Fisher criterion

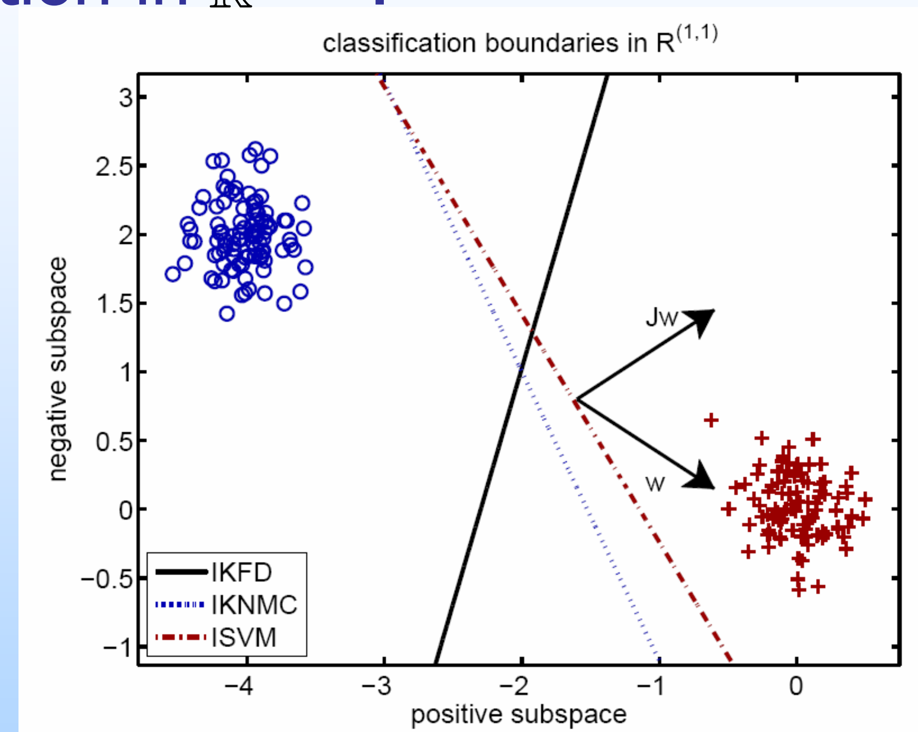
$$J(\mathbf{w}) := \frac{\left\langle \mathbf{w}, \Sigma^B_{\mathrm{pE}} \mathbf{w} \right\rangle_{\mathrm{pE}}}{\left\langle \mathbf{w}, \Sigma^W_{\mathrm{pE}} \mathbf{w} \right\rangle_{\mathrm{pE}}}$$

- Fisher Discriminant

$$f(\mathbf{z}) = \langle \mathbf{w}, \mathbf{z} \rangle_{\mathrm{pE}} + b \qquad b = -\frac{1}{2} \left\langle \boldsymbol{\mu}_+ + \boldsymbol{\mu}_-, \mathbf{w} \right\rangle_{\mathrm{pE}}$$

Geometrical Interpretation

- Illustration in $\mathbb{R}^{(1,1)}$:



- pE-FD is a linear classifier, intuitive decision boundary
- ISVM, IKNMC suffer from „reflection" with $\mathbf{J}$
- pE-FD identical to FD in Associated Euclidean space $\mathbb{R}^{p+q}$
- No kernel matrix preprocessing necessary!!

Indefinite Kernel Fisher Discriminant

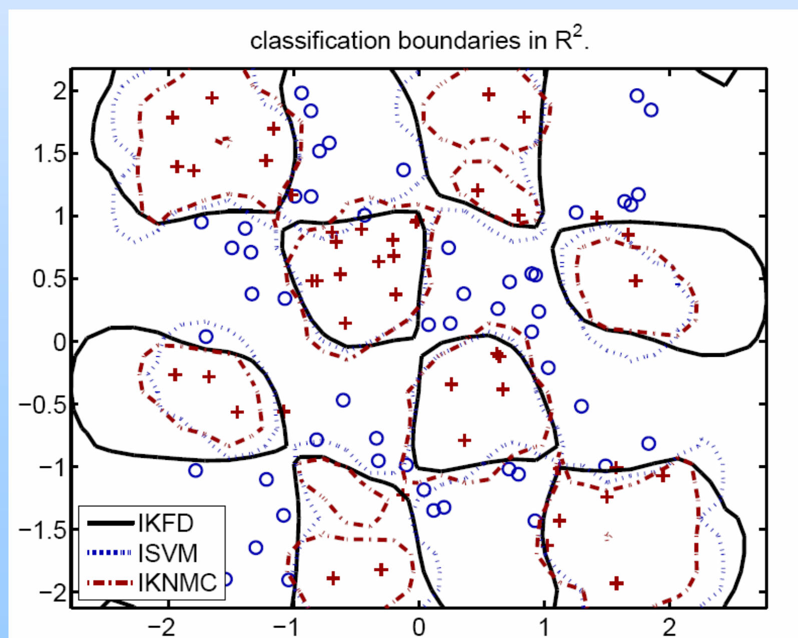
- Kernelization
  - Normal
    $$\mathbf{w} = \sum_{i=1}^{n} \alpha_i \phi(x_i)$$
  - Between-class scatter
    $$\left\langle \mathbf{w}, \Sigma_{\mathrm{pE}}^{B} \mathbf{w} \right\rangle_{\mathrm{pE}} = \boldsymbol{\alpha}^T \mathbf{K}(\mathbf{c}_+ - \mathbf{c}_-)(\mathbf{c}_+ - \mathbf{c}_-)^T \mathbf{K} \boldsymbol{\alpha}$$
  - Within-class scatter
    $$\left\langle \mathbf{w}, \Sigma_{\mathrm{pE}}^{W} \mathbf{w} \right\rangle_{\mathrm{pE}} = \boldsymbol{\alpha}^T \left(\mathbf{K}_+ \mathbf{H}_+ \mathbf{K}_+^T + \mathbf{K}_- \mathbf{H}_- \mathbf{K}_-^T \right) \boldsymbol{\alpha}$$
  - Maximization of regularized Fisher Criterion
    $$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T \mathbf{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \mathbf{N}_\beta \boldsymbol{\alpha}} \qquad \boldsymbol{\alpha} = \mathbf{N}_\beta^{-1} \mathbf{K}(\mathbf{c}_+ - \mathbf{c}_-)$$
- Indefinite KFD:
  $$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x) + b \qquad b = -\frac{1}{2} \boldsymbol{\alpha}^T \left(\frac{1}{n_+} \mathbf{K}_+ \mathbf{1}_{n_+} + \frac{1}{n_-} \mathbf{K}_- \mathbf{1}_{n_-} \right)$$
- ⇨ Correspondence to KFD with indefinite kernel matrix

Experiments 2D

- MATLAB Toolbox PRTools [http://prtools.org]
- Checkerboard dataset
  - 50+50 training samples
  - Indefiniteness by reflection-invariance:

$$\tau(x) := -x \qquad k(x, x') := \max(k_{\mathrm{rbf}}(x, x'), k_{\mathrm{rbf}}(x, \tau(x')))$$

  - Model selection by 10-fold CV for β, C, σ



Perfect point symmetry

Experiments 2D

- Quantitative aspects
  - Negative variance ratio $r := \left(\sum_{\lambda_i<0} |\lambda_i|\right) / \left(\sum_{\lambda_i} |\lambda_i|\right)$
  - Test errors over 500+500 samples
- Overall recognition accuracy
  - Cross-validated classifiers: IKFD lowest test error
- Fixed kernel parameter σ, 10-fold CV for C, β
  - ISVM good for weak indefiniteness
  - IKFD good for substantial indefiniteness

σ	$r\ (p,q)$	IKFD (β)	ISVM (C)	IKNMC
0.010	0.000 (98,2)	0.336 (10)	0.323 (10)	0.340
0.050	0.022 (82,18)	0.145 (10)	0.134 (10)	0.173
0.100	0.055 (66,34)	0.121 (10^{-1})	0.121 (1)	0.201
0.500	0.125 (51,49)	0.083 (1)	0.168 (1)	0.384
1.000	0.132 (52,48)	0.091 (10^{-3})	0.418 (1)	0.486
5.000	0.107 (50,50)	0.132 (10^{-2})	0.480 (1)	0.497
10.00	0.062 (51,49)	0.159 (10^{-3})	0.373 (10^{2})	0.494

Experiments Real World

- Polygon dataset
  - 2-classes, polygons of 5/7 vertices
  - Mod. Hausdorff-distance kernel $k(x, x') := -d_{MH}(x, x')^{\gamma}$
  - 10-fold CV of kernel and regularization parameters
  - Results 10-fold averaged, 100 train/3900 test samples

γ	mean (p, q)	IKFD	ISVM	IKNMC
0.2	(99.0,1.0)	0.021±0.006	0.021±0.006	0.089±0.027
0.5	(99.0,1.0)	0.019±0.006	0.018±0.004	0.110±0.034
0.7	(98.9,1.1)	0.020±0.004	0.018±0.004	0.118±0.037
1.0	(85.9,14.1)	0.019±0.009	0.029±0.007	0.129±0.041
2.0	(48.9,51.1)	0.017±0.008	0.094±0.057	0.152±0.051
5.0	(44.8,55.2)	0.102±0.021	0.131±0.030	0.218±0.081
7.0	(47.4,52.6)	0.111±0.027	0.237±0.058	0.253±0.093

- IKFD and ISVM outperform IKNMC
- IKFD better than ISVM for clearly indefinite data

Experiments Real World

- Chicken pieces dataset
  - 5-classes, 446 objects
  - nonsymmetric dissimilarity matrix (edit distance) [PHDSB06]
  - Symmetric distance kernel $k(x,x') := -((d(x,x') + d(x',x))/2)^2$
  - 10-fold CV of regularization parameters
  - Results 20-fold averaged, 75% train/25% test

Method	Test error
IKFD	0.0785 ± 0.0340
ISVM	0.1029 ± 0.0273
IKNMC	0.2966 ± 0.0496

- IKFD lower test-error than ISVM and IKNMC

Indefinite Kernel Discriminant Feature Extraction

Indefinite Kernel Mahalanobis Distance

- Covariance operator in indefinite spaces

$$Cv := \frac{1}{n}\sum_{i=1}^{n} \tilde{\phi}(x_i) \left\langle \tilde{\phi}(x_i), v \right\rangle_{\mathcal{K}} = \frac{1}{n}\tilde{\Phi}\tilde{\Phi}^T J v = C^{|\mathcal{K}|} J v$$

  - is pd in Krein-sense $\langle \psi, C\psi \rangle_{\mathcal{K}} \geq 0$
- Indefinite Kernel Mahalanobis, **I**nvertible **C**ovariance
  - direct extension of pd case:

$$d_{IC}^2(x) := \left\langle \tilde{\phi}(x), C^{-1}\tilde{\phi}(x) \right\rangle_{\mathcal{K}} = n(\tilde{\mathbf{k}}_x)^T (\tilde{K}^-)^2 \tilde{\mathbf{k}}_x$$

- Application classwise yields feature vector:

$$f_{IKM-IC}(x) := \left(d_{IC}^{[1]}(x), \ldots, d_{IC}^{[c]}(x) \right)^T \in \mathbb{R}^c$$

Indefinite Kernel Mahalanobis Distance

- Indefinite Kernel Mahalanobis, **F**ull **K**ernel
  - Use of inter-class information by KPCA
  - Direct extension of pd case

$$(d_{FK}^{[j]}(x))^2 := \frac{n^{[j]}}{2}(\tilde{\mathbf{k}}_x^{[j]})^T(\tilde{K}_{reg}^{[j]})^{-1}\tilde{\mathbf{k}}_x^{[j]}$$

with

$$\tilde{\mathbf{k}}_x^{[j]} := \mathbf{k}_x - \frac{1}{n^{[j]}}K^{[j]}\mathbf{1}_{n^{[j]}}$$

$$K_{reg}^{[j]} := \tilde{K}^{[j]} + \alpha_j I_n$$

$$\tilde{K}^{[j]} := K^{[j]}H^{[j]}K^{[j]T} \in \mathbb{R}^{n\times n}$$

- Feature vector

$$f_{IKM-FK}(x) := \left(d_{FK}^{[1]}(x), \ldots, d_{FK}^{[c]}(x)\right)^T \in \mathbb{R}^c$$

Indefinite Kernel Fisher Disciminant Features

- Pd-case: Generalized discriminant analysis [BA00]
- Similar to IKFD, now multi-class setting
- Search $W = [w_1, \ldots, w_{c-1}] \in \mathcal{K}^{c-1}$ maximizing

$$J(W) := \frac{\det(\langle W, \Sigma_B W\rangle_{\mathcal{K}})}{\det(\langle W, \Sigma_W W\rangle_{\mathcal{K}})}$$

- Solved by computing $W = \Phi\boldsymbol{\alpha}$ with

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{c-1}) \in \mathbb{R}^{n\times(c-1)}$$

  and solving eigenvalue problem

$$(N_\beta^{-1} M)\boldsymbol{\alpha}_j = \lambda_j \boldsymbol{\alpha}_j$$

- Projection onto eigenvectors yield features:

$$f_{IKF}(x) := (\langle w_1, \phi(x)\rangle_{\mathcal{K}}, \ldots, \langle w_{c-1}, \phi(x)\rangle_{\mathcal{K}})^T = \boldsymbol{\alpha}^T \mathbf{k}_x \in \mathbb{R}^{c-1}$$

Experiments

- Indefinite multiclass datasets
  - Negative variance ratio $r_{neg} := \left(\sum_{\lambda_i<0} |\lambda_i|\right) / \left(\sum_{\lambda_i} |\lambda_i|\right)$
  - Hold out ratio β
  - Kernel $k = s$ or $k = -d^2$, centered

	Dissimilarity	Kernel	$c\ (n^{[j]})$	β	$r_{\mathrm{neg}}(p, q)$
Cat-cortex	Prior knowl.	$-d^2$	4 (10–19)	0.80	0.19 (35, 18)
Protein	Evolutionary	$-d^2$	4 (30–77)	0.80	0.00 (167, 3)
News-COR	Correlation	$-d^2$	4 (102–203)	0.60	0.19 (127,208)
ProDom	Structural	s	4 (271–1051)	0.25	0.01 (518, 90)
Chicken29	Edit-dist.	$-d^2$	5 (61–117)	0.80	0.31 (192,166)
Files	Compression	$-d^2$	5 (60–255)	0.50	0.02 (392, 63)
Pen-ANG	Edit-dist.	$-d^2$	10 (334–363)	0.15	0.24 (261,269)
Zongker	Shape-match.	s	10 (200)	0.25	0.36 (274,226)

(average over 20 hold out drawings, centered

Experiments

- Feature/classifier settings:
  - Features: IKM-IC, IKM-FK, IKF
  - Classifiers in c / (c-1) dim space: Nearest Mean (NM), Fisher Discriminant (FD), Quadratic Discriminant (QD), k-nearest-neighbour (KNN)
  - Indefinite Kernel Classifiers as Reference: Kernel Fisher Discriminant (IKFD), Support-Vector-Machine (SVM), Kernel-k-Nearest-Neighbour (IKNN)
  - 10-fold Cross validation of regularization parameters

Experiments

- Recognition results
  - average (std) test error over 25 hold out runs

Classifier+Features	Cat-cortex	Protein	News-COR	ProDom
NM+IKM-IC	45.5 (13.2)	21.2 (7.7)	38.6 (2.4)	15.0 (3.5)
NM+IKM-FK	10.9 (6.3)	2.1 (2.6)	24.9 (2.5)	6.4 (2.4)
NM+IKF	12.6 (5.7)	0.1 (0.4)	24.1 (1.8)	2.0 (0.6)
FD+IKM-IC	42.2 (11.3)	25.9 (5.5)	39.7 (2.6)	9.4 (3.0)
FD+IKM-FK	10.3 (5.4)	1.1 (2.0)	24.2 (2.0)	1.7 (0.6)
FD+IKF	11.2 (5.2)	0.2 (0.5)	24.2 (3.1)	1.6 (0.6)
QD+IKM-IC	48.5 (12.4)	11.9 (4.5)	41.4 (3.0)	3.6 (0.9)
QD+IKM-FK	22.7 (6.7)	0.5 (0.8)	25.5 (2.7)	2.0 (0.7)
QD+IKF	18.4 (7.4)	0.5 (1.3)	24.4 (3.1)	1.5 (0.5)
KNN+IKM-IC	43.9 (8.6)	19.8 (7.4)	42.9 (3.1)	5.0 (1.6)
KNN+IKM-FK	11.3 (6.5)	0.6 (1.7)	25.7 (1.7)	2.2 (0.9)
KNN+IKF	11.7 (6.5)	0.2 (0.5)	24.7 (2.2)	1.6 (0.7)
IKFD	10.6 (5.6)	0.3 (0.7)	23.6 (2.4)	2.0 (0.6)
ISVM	16.5 (5.7)	0.5 (0.8)	24.4 (2.3)	1.6 (0.6)
IKNN	15.6 (5.8)	4.7 (5.2)	29.6 (2.3)	3.1 (0.8)

Experiments

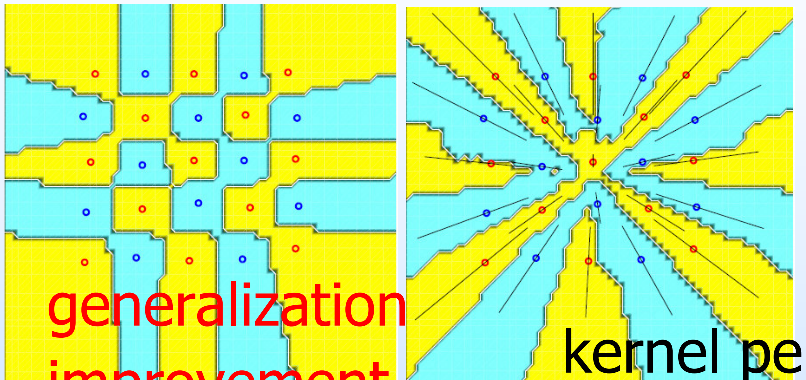
- Findings:
  - High std.-dev, caution of overinterpretation
  - IKM-IC performs worse than other features
    - ⇒ Assumption of IC may be wrong
    - ⇒ Between-class information is ignored
  - IKF mostly preferable over IKM-IC/FK features
  - KNN-classifier best on features ⇒ nonlinear classifiers beneficial
  - Features yield results in the range of the reference classifiers
  - Reference Classifiers: IKFD mostly better than ISVM or IKNN

Summary and Conclusions

- Kernel Discriminant Classifiers
  - KQDA-classification good for positive definite kernels
  - IKFD, ISVM: Applicable to indefinite kernels, sound geometrical interpretation
  - IKFD: Superiority over ISVM, IKNMC on 2D and benchmark data
- Indefinite Kernel Discriminant Feature Extraction
  - IKF, IKM-FK allow reference classification performance
  - IKF, IKFD are identical to their positive definite counterpart, no data „Euclideanization“ required.
- Indefinite Kernel Methods
  - Indefinite kernels practically relevant: result from inclusion of prior knowledge, kernel combination, dissimilarities
  - Interpretation of indefinite kernels in Krein-spaces: basis for geometrical/numerical/statistical analysis and new methods

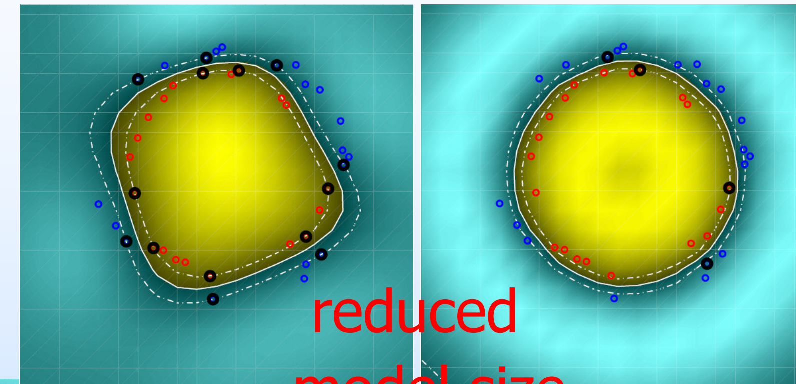
Application in general kernel methods:

kernel-nn-classification:



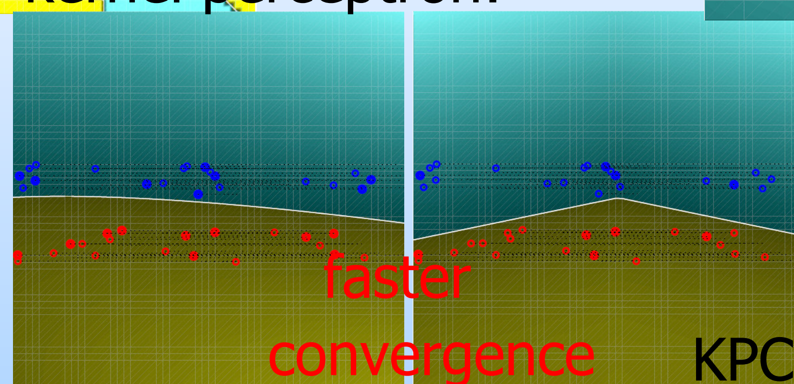
generalization
improvement

SVM:



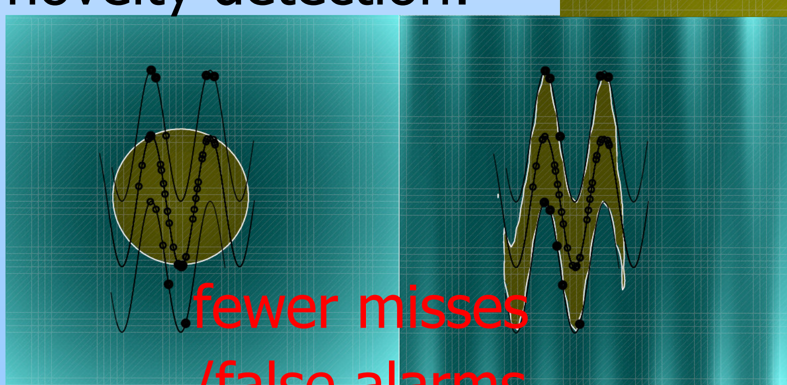
reduced
model size

kernel perceptron:



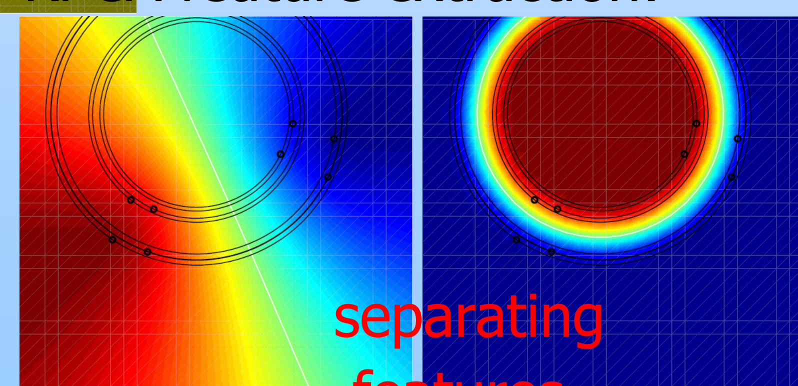
faster
convergence

novelty detection:



fewer misses
/false alarms

KPCA feature extraction:



separating
features

Thank You!

Questions?

References

- [BA00] G. Baudat and F. Anouar: „Generalized Discriminant Analysis using a kernel approach", Neural Computation, 12(10):2385-2404,2000
- [DHS01] R. Duda, P. Hart and D. Stork, Pattern Classification, 2nd ed. John Wiley & Sons, Inc., 2001.
- [DJRPPT04] R. Duin, O. Juszcak, D. de Ridder, P. Paclik, E. Pekalska, and D. Tax, „PR-Tools",2004, http://prtools.org.
- [G85] L. Goldfarb. „A new approach to pattern recognition", Progress in Pattern Recongition 2, pp. 241-402, Elsevier, 1985.
- [GHBO99] T. Graepel, R. Herbrich, P. Bollmann-Sdorra,K. Obermayer. „Classification on pariwise proximity data", NIPS 11, pp. 438-444, MIT Press, 1999.
- [Ha05] B. Haasdonk. „Feature space interpretation of SVMs with indefinite kernels", IEEE TPAMI, 27(4):482-492, 2005.
- [HP08] B. Haasdonk, E. Pekalska: „Classification with kernel Mahalanobis distances", In Proc. GfKl 2008.
- [HP08b] B. Haasdonk, E. Pekalska: „Indefinite Kernel Fisher Discriminant", In Proc. ICPR 2008.
- [HP10] B. Haasdonk, E. Pekalska: „Indefinite Kernel Discriminant Analysis", In Proc. COMPSTAT 2010.
- [MRWSM99] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, K.R. Müller:"Fisher Discriminant Analysis with Kernels", In Neural Networks for Signal Processing, pp 41-48
- [OMCS04] C.S. Ong, X. Mary, S. Canu, A.J. Smola: „Learning with nonpositive kernels", In ICML, pp 639—646, ACM Press, 2004
- [PD05] E. Pekalska, and R.P.W. Duin: „The dissimilarity representation for pattern recognition, Foundations and Applications". World Scientific, Singapore 2005
- [PH09] E. Pekalska and B. Haasdonk, „Kernel Quadratic Discriminant Analysis with Positive Definite and Indefinite Kernels", TPAMI, 2009.
- [PPD01] E. Pekalska, P. Paclik, R. Duin. „A generalized kernel approach to dissimilarity based classification" Journal of Machine Learning Research, 2:175-211, 2001.
- [ROM98] G. Rätsch, T. Onoda, K.-R. Müller. "Soft margins for Adaboost", TR NC-TR-1998-021, Royal Holloway College, University of London, UK, 1998.
- [SC04] J. Shawe-Taylor and N. Cristianini, „Kernel Methods for Pattern Analysis", Cambridge University Press, 2004
- [SS02] B. Schölkopf and A. Smola, „Learning with Kernels", MIT-Press, 2002.