Introduction to Kernel Methods

Bernhard Schölkopf

Max Planck Institute for Biological Cybernetics
72076 Tübingen, Germany

www.kyb.tuebingen.mpg.de/~bs

Acknowledgement

Olivier Bousquet
Olivier Chapelle
André Elisseeff
Ulrike von Luxburg
Jason Weston

Learning Problem

Suppose we are given data

$$(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathcal{Y}$$

where $(x_i, y_i) \sim \mathrm{P}(x, y)$.
We want to estimate a function $f : \mathcal{X} \to \mathcal{Y}$ such that

$$R[f] = \int_{\mathcal{X}} l(f(x), y)\, d\mathrm{P}(x, y)$$

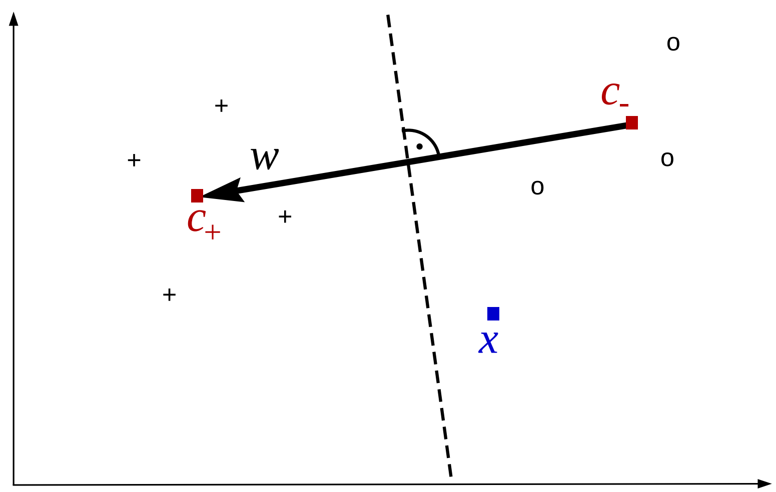
is minimized.
Here, $l(f(x), y)$ is the loss incurred when predicting $f(x)$ if the true output is y.

Special case: $\mathcal{Y} = \{\pm 1\}, l(f(x), z) = \frac{1}{2}|f(x) - y|$:
binary pattern recognition

An Example of a Pattern Recognition Algorithm

Idea: classify points x according to which of the two class means is closer.

$$c_+ := \frac{1}{m_+} \sum_{y_i=1} x_i, \quad c_- := \frac{1}{m_-} \sum_{y_i=-1} x_i$$



- Decision function: hyperplane with normal vector $w := c_+ - c_-$
- How about problems that are not linearly separable?

Kernel Feature Spaces

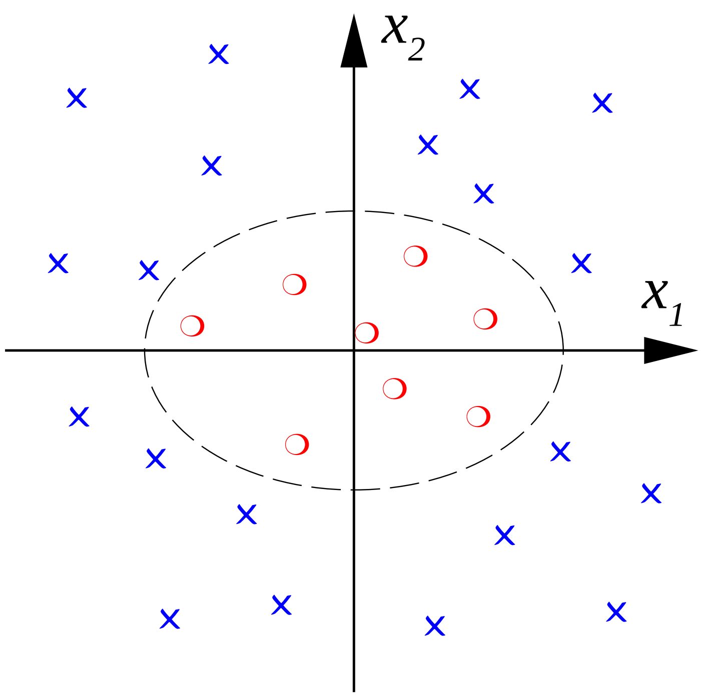
Preprocess the inputs with

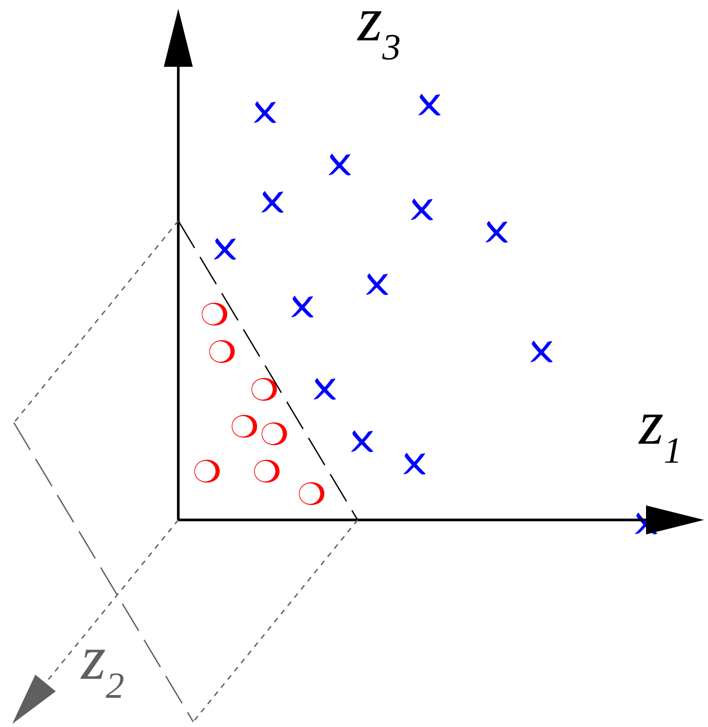
$$\begin{aligned}\Phi : \mathcal{X} &\to \mathcal{H} \\ x &\mapsto \Phi(x),\end{aligned}$$

where $\mathcal{H}$ is a dot product space, and learn the mapping from $\Phi(x)$ to y.

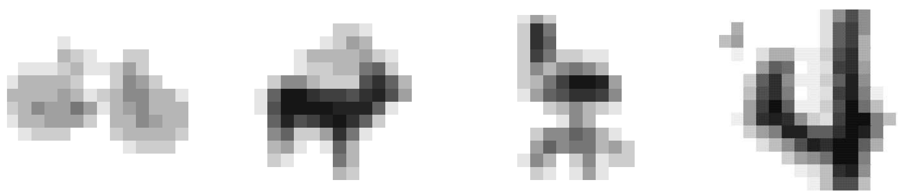
Example: All Degree 2 Monomials

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)$$





General Product Feature Space



How about patterns $x \in \mathbb{R}^N$ and product features of order d?

Here, $\dim(\mathcal{H})$ grows like N^d.

E.g. $N = 16 \times 16$, and $d = 5 \longrightarrow$ dimension 10^{10}

The Kernel Trick, $N = d = 2$

$$\begin{aligned}
\left\langle \Phi(x), \Phi(x') \right\rangle &= (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)(x'^2_1, \sqrt{2}\, x'_1 x'_2, x'^2_2)^\top \\
&= (x_1 x'_1 + x_2 x'_2)^2 \\
&= \left\langle x, x' \right\rangle^2 \\
&= : k(x, x')
\end{aligned}$$

$\longrightarrow$ the dot product in $\mathcal{H}$ can be computed from the dot product in $\mathbb{R}^2$

The Kernel Trick, $N = d = 2$

$$\begin{aligned}
\left\langle \Phi(x), \Phi(x') \right\rangle &= (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)(x'^2_1, \sqrt{2}\, x'_1 x'_2, x'^2_2)^\top \\
&= (x_1 x'_1 + x_2 x'_2)^2 \\
&= \left\langle x, x' \right\rangle^2 \\
&= : k(x, x')
\end{aligned}$$

$\longrightarrow$ the dot product in $\mathcal{H}$ can be computed from the dot product in $\mathbb{R}^2$

More generally: for $x, x' \in \mathbb{R}^N$, $d \in \mathbb{N}$,

$$\langle x, x' \rangle^d = \left(\sum_{j=1}^{N} x_j \cdot x'_j \right)^d = \sum_{j_1, \ldots, j_d = 1}^{N} x_{j_1} \cdot \cdots \cdot x_{j_d} \cdot x'_{j_1} \cdot \cdots \cdot x'_{j_d} = \langle \Phi(x), \Phi(x') \rangle \, .$$

Positive Definite Kernels

Let $\mathcal{X}$ be a nonempty set. The following two are equivalent:

- k is *positive definite (pd)*, i.e., k is symmetric, and for
  - any set of training points $x_1, \ldots, x_m \in \mathcal{X}$ and
  - any $a_1, \ldots, a_m \in \mathbb{R}$

  we have

$$\sum_{i,j} a_i a_j K_{ij} \geq 0, \quad \text{where } K_{ij} := k(x_i, x_j)$$

- there exists a map Φ into a dot product space $\mathcal{H}$ such that

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

$\mathcal{H}$ is a so-called *reproducing kernel Hilbert space.*

Special case of positive definite kernels: "Mercer kernels"

The Kernel Trick — Summary

- *any* algorithm that only depends on dot products can benefit from the kernel trick
- $\mathcal{X}$ need not be a vector space
- think of the kernel as a (nonlinear) *similarity measure*
- examples of common kernels:

$$\text{Polynomial} \quad k(x, x') = (\langle x, x' \rangle + c)^d$$

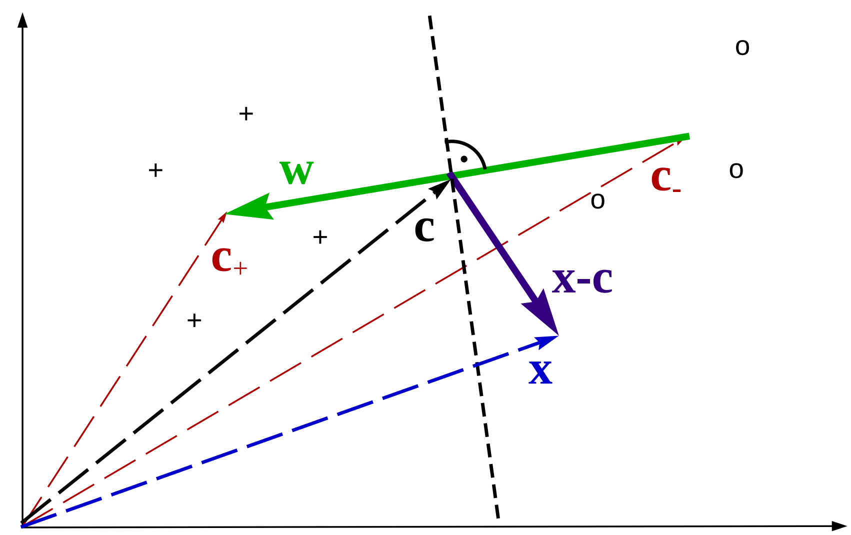
$$\text{Gaussian} \quad k(x, x') = \exp(-\|x - x'\|^2 / (2\,\sigma^2))$$

- Kernels are studied also in approximation theory *(Micchelli, 1986; Wahba, 1990; Berg et al., 1984)* and in the Gaussian Process prediction community (covariance functions) *(Weinert, 1982; Wahba, 1990; Williams, 1998; MacKay, 1998)*

An Example of a Kernel Algorithm

Classify points $\mathbf{x} := \Phi(x)$ *in feature space* according to which of the two class means is closer.

$$\mathbf{c}_+ := \frac{1}{m_+} \sum_{\{i:y_i=1\}} \Phi(x_i), \quad \mathbf{c}_- := \frac{1}{m_-} \sum_{\{i:y_i=-1\}} \Phi(x_i)$$



Compute the sign of the dot product between $\mathbf{w} := \mathbf{c}_+ - \mathbf{c}_-$ and $\mathbf{x} - \mathbf{c}$.

An Example of a Kernel Algorithm, ctd.

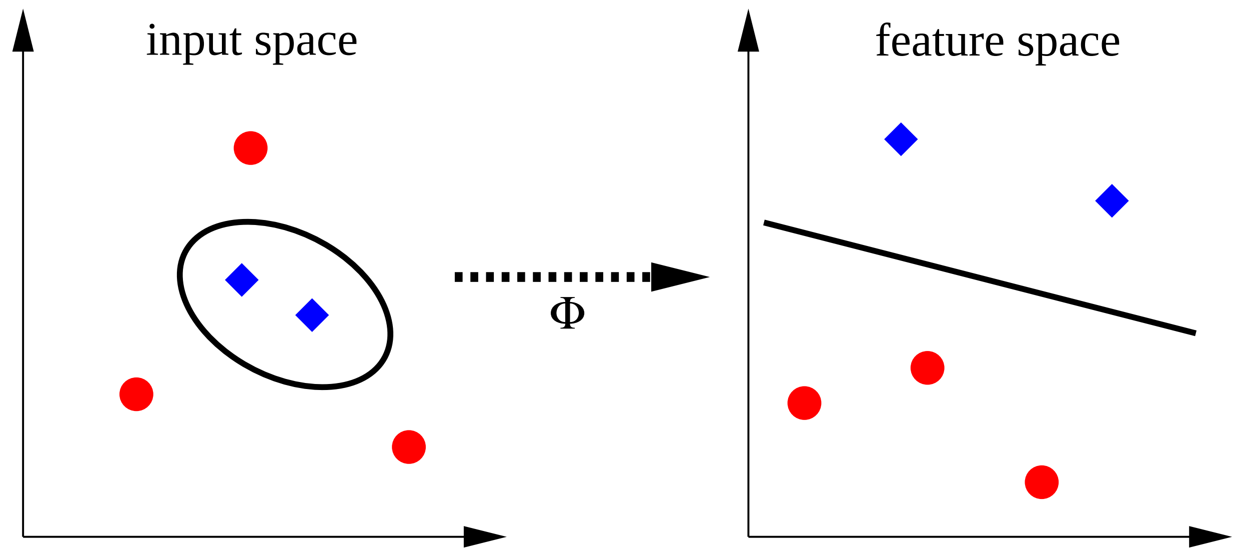
$$
\begin{aligned}
f(x) &= \operatorname{sgn}\left(\frac{1}{m_+}\sum_{\{i:y_i=1\}}\langle\Phi(x),\Phi(x_i)\rangle - \frac{1}{m_-}\sum_{\{i:y_i=-1\}}\langle\Phi(x),\Phi(x_i)\rangle + b\right)\\
&= \operatorname{sgn}\left(\frac{1}{m_+}\sum_{\{i:y_i=1\}} k(x,x_i) - \frac{1}{m_-}\sum_{\{i:y_i=-1\}} k(x,x_i) + b\right)
\end{aligned}
$$

with the constant offset

$$
b = \frac{1}{2}\left(\frac{1}{m_-^2}\sum_{\{(i,j):y_i=y_j=-1\}} k(x_i,x_j) - \frac{1}{m_+^2}\sum_{\{(i,j):y_i=y_j=1\}} k(x_i,x_j)\right).
$$

- if k is a density: Parzen windows interpretation

Support Vector Classifiers



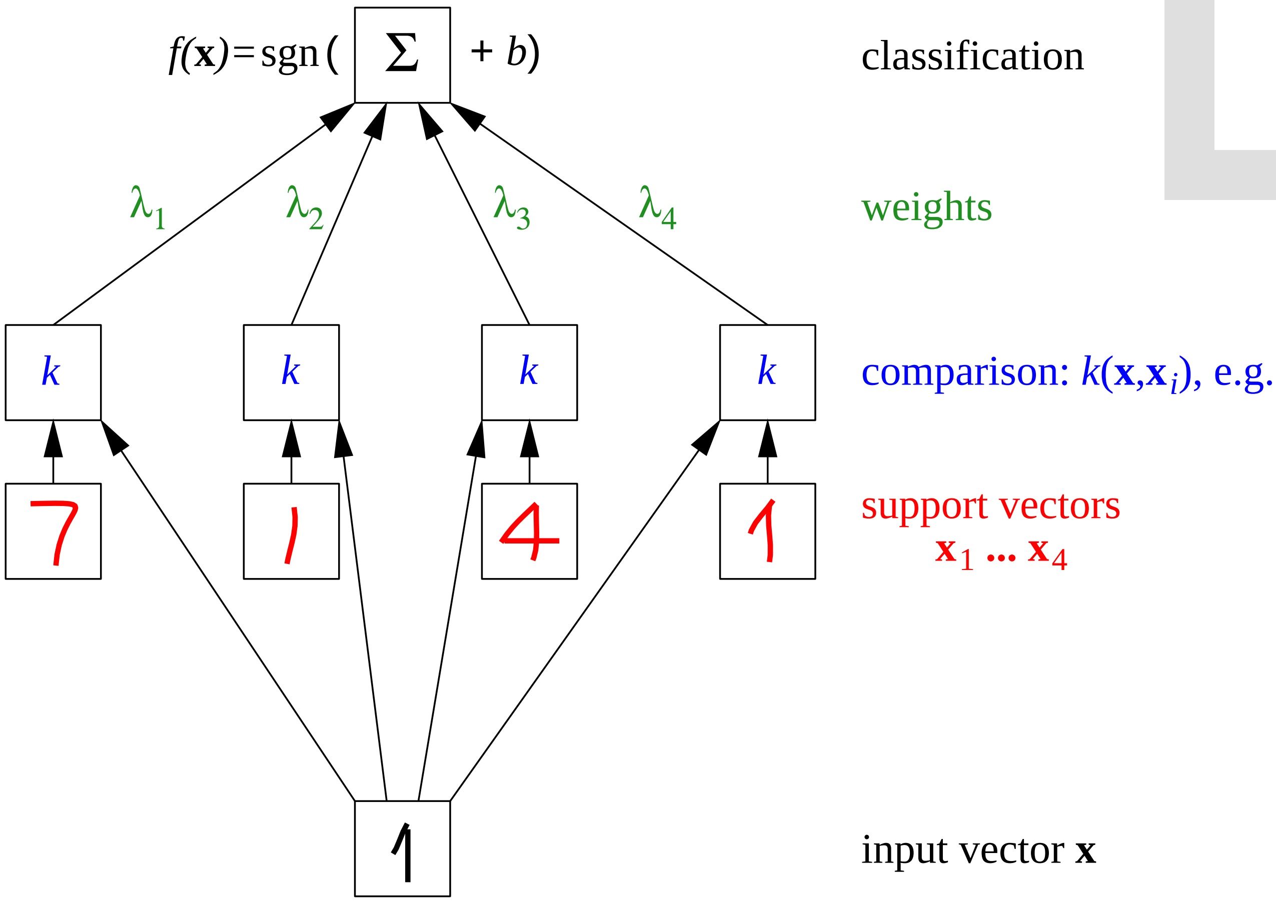
- large margin separation in $\mathcal{H}$
- sparse expansion of solution in terms of SVs:

$$f(x) = \mathrm{sgn}(\sum_i \lambda_i k(x_i, x) + b)$$

- unique solution found by convex QP
- Demo

Boser, Guyon, and Vapnik (1992)

The SVM Architecture



$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum \lambda_i \cdot k(\mathbf{x}, \mathbf{x}_i) + b \right)$$

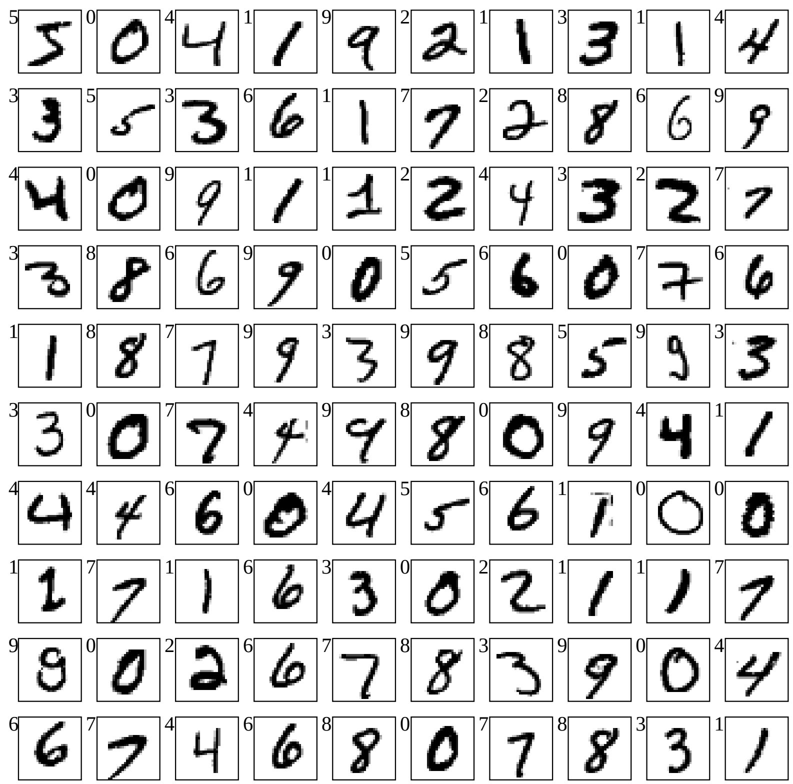
$k(\mathbf{x},\mathbf{x}_i) = (\mathbf{x} \cdot \mathbf{x}_i)^d$

$k(\mathbf{x},\mathbf{x}_i) = \exp(-\|\mathbf{x} - \mathbf{x}_i\|^2 / c)$

$k(\mathbf{x},\mathbf{x}_i) = \tanh(\kappa(\mathbf{x} \cdot \mathbf{x}_i) + \theta)$

MNIST Benchmark

handwritten character benchmark (60000 training & 10000 test examples, 28×28)



MNIST Error Rates

Classifier	test error	reference
linear classifier	8.4%	*Bottou et al. (1994)*
3-nearest-neighbour	2.4%	*Bottou et al. (1994)*
SVM	1.4%	*Burges and Schölkopf (1997)*
Tangent distance	1.1%	*Simard et al. (1993)*
LeNet4	1.1%	*LeCun et al. (1998)*
Boosted LeNet4	0.7%	*LeCun et al. (1998)*
Translation invariant SVM	0.56%	*DeCoste and Schölkopf (2002)*

Dimensionality of the Feature Space

Note: the SVM system that holds the record on the MNIST set used a polynomial kernel of degree 9, corresponding to a feature space of dimensionality $\approx 3.2 \cdot 10^{20}$.

"Curse of Dimensionality"?

Statistical Learning Theory: there is a curse of *capacity*, not of *dimensionality*

Pattern Recognition

Learn $f : \mathcal{X} \to \{\pm 1\}$ from examples

$(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \{\pm 1\}$, each pair generated from $\mathrm{P}(x, y)$, such that the expected misclassification error on a test set, also drawn from $\mathrm{P}(x, y)$,

$$R[f] = \int \frac{1}{2} |f(x) - y)| \, d\mathrm{P}(x, y),$$

is minimal *(Risk Minimization (RM))*.

Problem: P is unknown. $\longrightarrow$ need an *induction principle*.

Empirical risk minimization (ERM): replace the average over $\mathrm{P}(x, y)$ by an average over the training sample, i.e. minimize the training error

$$R_{\mathrm{emp}}[f] = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} |f(x_i) - y_i|$$

Convergence of Means to Expectations

Law of large numbers: for every $f \in \mathcal{F}$,

$$\lim_{m \to \infty} \mathrm{P}\{|R[f] - R_{\mathrm{emp}}[f]| > \epsilon\} = 0$$

for all $\epsilon > 0$.

Does this imply that ERM will give us the optimal result in the limit of infinite sample size (*"consistency"* of empirical risk minimization)?

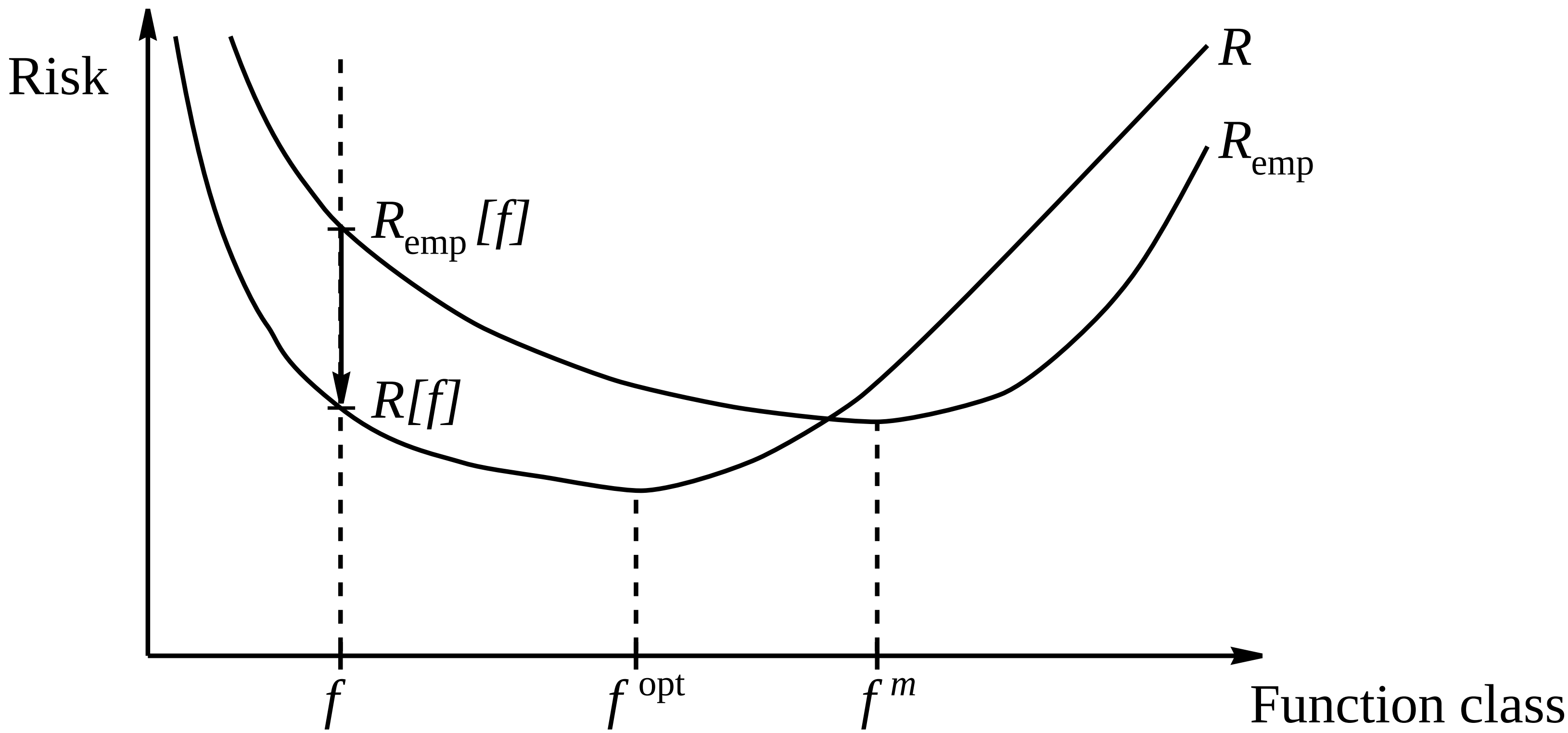
No.

Vapnik and Chervonenkis showed that ERM is (nontrivially) consistent if and only if the convergence is uniform:

$$\lim_{m \to \infty} \mathrm{P}\{\sup_{f \in \mathcal{F}} (R[f] - R_{\mathrm{emp}}[f]) > \epsilon\} = 0$$

for all $\epsilon > 0$.

Consistency and Uniform Convergence



How about taking $\mathcal{F} = \{\text{all functions mapping } \mathcal{X} \text{ to } \{\pm 1\}\}$?
Fix m. For every "good" function there exists a "bad" function with the same value of R_{emp}, and possibly rather different R.

Capacity

Vapnik, Chervonenkis and others give conditions for uniform convergence in terms of capacity concepts of the function class, e.g.

- the VC-entropy grows sublinearly with m
- the VC-dimension is finite
- the entropy numbers are well-behaved

(e.g. Vapnik and Chervonenkis, 1974; Vapnik, 1998; Shawe-Taylor et al., 1998; Williamson et al., 1998; Alon et al., 1997)

To have a low test error, we need a low training error and low capacity.

Justifications for Large Margins

- VC-dimension: $h \leq R^2/\rho^2$, where ρ is the margin and R is the radius of the smallest sphere containing the data *(Vapnik, 1979)*
- fat-shattering dimension and data dependent SRM *(Gurvits, 1997; Shawe-Taylor et al., 1998)*
- regularization theory *(Girosi, 1998; Smola and Schölkopf, 1998)* and Bayesian MAP estimation *(Kimeldorf and Wahba, 1970; Poggio and Girosi, 1990)*
- algorithmic stability *(Bousquet and Elisseeff, 2001)*
- Rademacher averages *(Kolchinskiĭ et al., 2001; Mendelson, 2001; Bousquet, 2002)*
- compression/MDL *(von Luxburg et al., 2002)*

Compression Bound

Given: a finite function class $\mathcal{F}$.

Denote by C the *compression coefficient* of the training labels given the training inputs, using functions from $\mathcal{F}$.

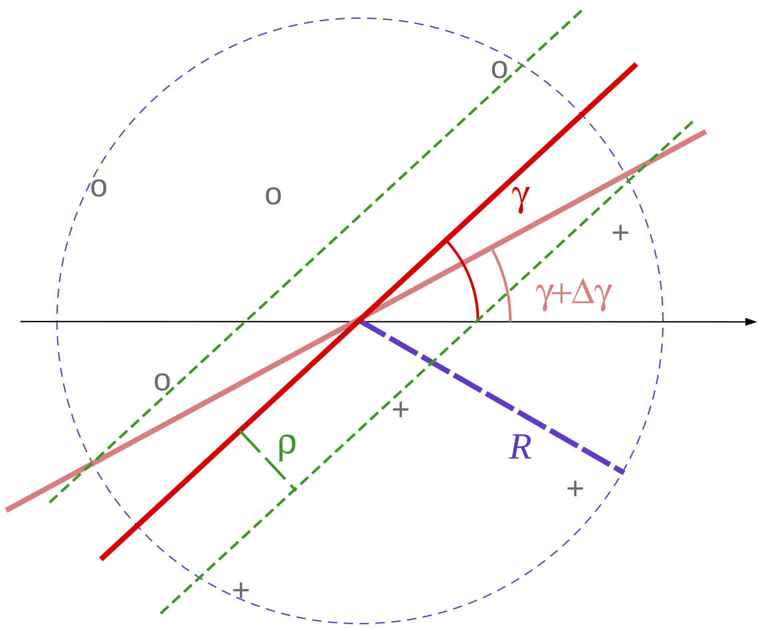
Sender		Receiver
$x_1, \ldots, x_m$		$x_1, \ldots, x_m$
$y_1, \ldots, y_m$		
f	$\longrightarrow$	$f(x_1), \ldots, f(x_m)$

Theorem. For all $f \in \mathcal{F}$, with probability $\geq 1 - \delta$,

$$R[f] \leq 2 \log(2) C - \frac{\ln(\delta)}{m}.$$

(Vapnik (1995), cf. also Littlestone and Warmuth (1986))

Maximum Margin vs. MDL — 2D Case

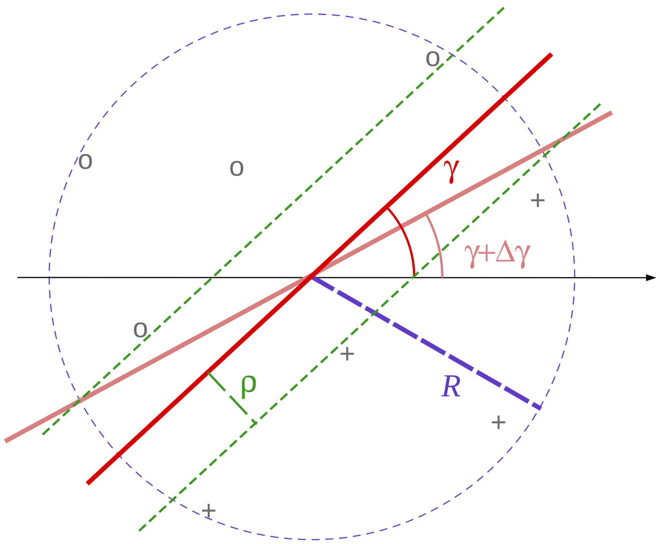


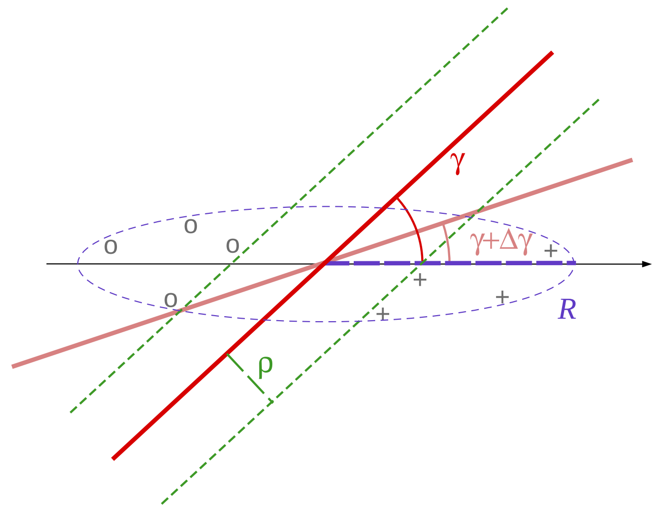
Can perturb γ by $\Delta\gamma$ with $|\Delta\gamma| < \arcsin\frac{\rho}{R}$ and still correctly separate the data.

Hence only need to transmit γ with accuracy $\Delta\gamma$ *(von Luxburg et al., 2002)*.

Can be done by computing a $\Delta\gamma$-cover of $\mathcal{F} = \{\text{hyperplanes with } \|w\| = 1\}$

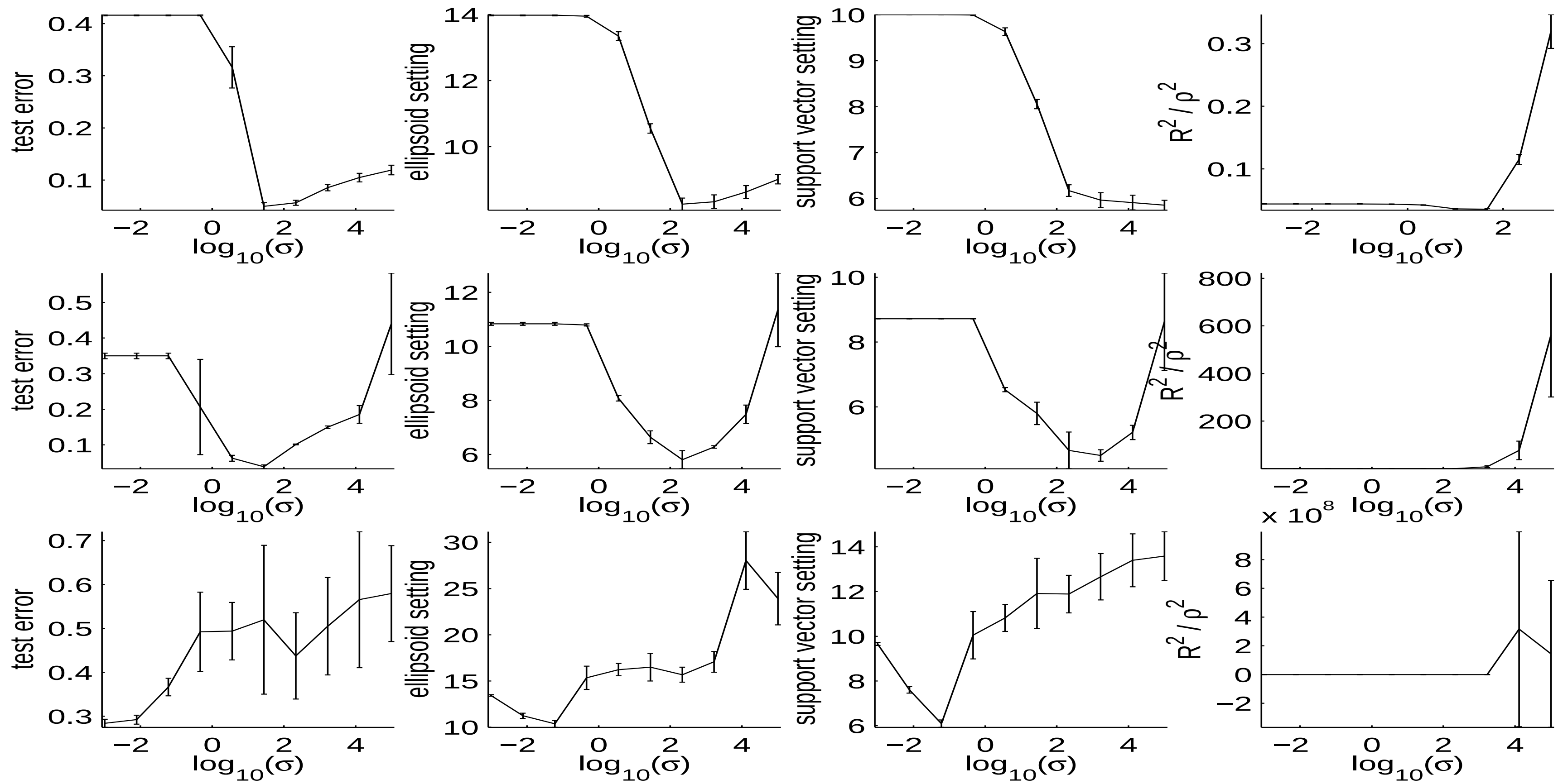
Ellipsoid Case





- ellipsoid setting: different directions imply different $\Delta\gamma$
- axes of ellipsoid can be computed from kernel eigenvalues

Experiments: Selecting σ in a Gaussian Kernel



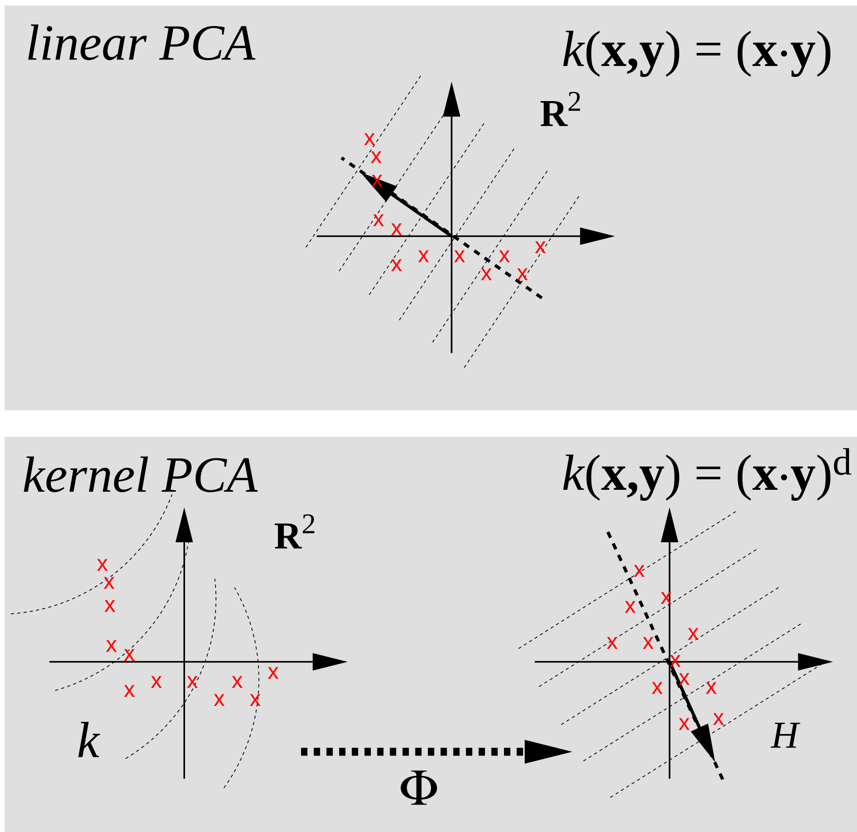
Datasets: USPS ($m = 500$)
Wisconsin breast cancer ($m = 200$)
Abalone ($m = 500$)

Further Kernel Algorithms — Design Principles

1. "Kernel module"
   - similarity measure $k(x, x')$, where $x, x' \in \mathcal{X}$
   - data representation
     (in associated feature space where $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$)
     — thus can construct geometric algorithms
   - function class ("representer theorem," $f(x) = \sum_i \alpha_i k(x, x_i)$)
2. "Learning module"
   - classification
   - quantile estimation / novelty detection
   - feature extraction
   - ...

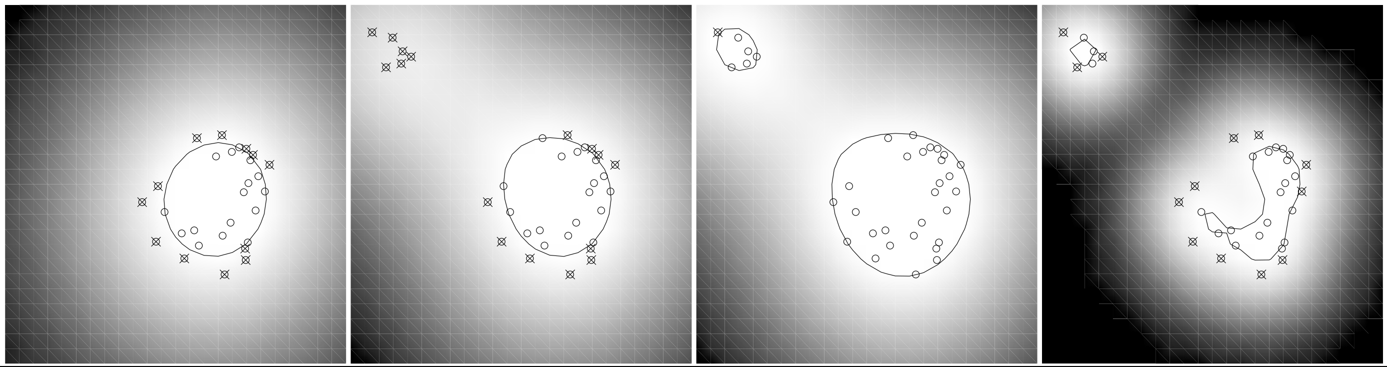
Kernel PCA

(Schölkopf et al., 1998)



Demo ○

One-Class SVMs



ν, width c	0.5, 0.5	0.5, 0.5	0.1, 0.5	0.5, 0.1
SVs/OLs	0.54, 0.43	0.59, 0.47	0.24, 0.03	0.65, 0.38

(using $k(x, y) = \exp(-\frac{\|x-y\|^2}{c})$)

- Jet engine condition monitoring *(Hayton et al., 2001)*
- Network intrusion detection *(Wankadia et al., 2001)*
- Outlier detection *(Schölkopf et al., 2001)*
- Information retrieval

Kernel Dependency Estimation *(Weston et al., 2002)*

Given two sets $\mathcal{X}$ and $\mathcal{Y}$ with kernels k and k', and training data (x_i, y_i).

Estimate a dependency $\mathbf{w} : \mathcal{H} \to \mathcal{H}'$

$$\mathbf{w}(\cdot) = \sum_{ij} \alpha_{ij} \Phi'(y_j) \langle \Phi(x_i), \cdot \rangle \, .$$

This can be evaluated in various ways, e.g., given an x, we can compute the pre-image

$$y = \operatorname{argmin}_{\mathcal{Y}} \|\mathbf{w}(\Phi(x)) - \Phi'(y)\|.$$

A convenient way of learning the α_{ij} is to work in the kernel PCA basis.

Application to Image Completion

Shown are all digits where at least one of the two algorithms makes a mistake (73 mistakes for k-NN, 23 for KDE).

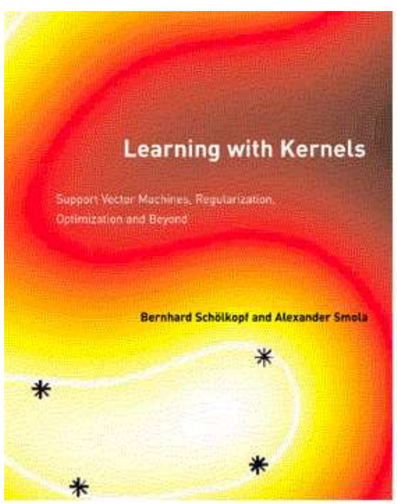
(from Weston et al. (2002))

Kernel Machines Research

- algorithms/tasks: KDE, feature selection *(Weston et al., 2002)*, multi-label-problems *(Elisseeff & Weston, 2001)*, unlabelled data *(Szummer & Jaakkola, 2002)*, ICA *(Harmeling et al., 2002)*, canonical correlations *(Bach & Jordan, 2002; Kuss, 2002)*
- optimization and implementation: QP, SDP *(Lanckriet et al., 2002)*, online versions, ...
- theory of empirical inference: sharper capacity measures and bounds *(Bartlett, Bousquet, & Mendelson, 2002)*, generalized evaluation spaces *(Mary & Canu, 2002)*, ...
- kernel design
  - transformation invariances *(Chapelle and Schölkopf, 2002)*
  - kernels for discrete objects *(Haussler, 1999; Watkins, 2000; Lodhi et al., 2000; Cristianini and Shawe-Taylor, 2000; Vert, 2002)*
  - kernels based on generative models *(Jaakkola and Haussler, 1999; Seeger, 1999; Tsuda et al., 2002)*
  - local kernels *(e.g., Zien et al., 2000)*
  - complex kernels from simple ones *(Haussler, 1999; Bartlett and Schölkopf, 2001)*, global kernels from local ones *(Kondor and Lafferty, 2002)*
  - functional calculus for kernel matrices *(Schölkopf et al., 2002)*
  - model selection, e.g., via alignment *(Cristianini et al., 2001)*

Conclusion

- crucial ingredients of SV algorithms: kernels that can be represented as dot products, and large margin regularizers
- kernels allow the formulation of a multitude of geometrical algorithms (Parzen windows, SV pattern recognition, SV quantile estimation, kernel PCA,...) that work very well in practice
- kernels unify three aspects of empirical inference: similarity measures, function classes, and data representations. The choice of a kernel is crucial, and it is not a problem of statistics.



For further information, cf.
`http://www.kernel-machines.org`,
`http://www.learning-with-kernels.org`

References

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

P. L. Bartlett and B. Schölkopf. Some kernels for structured data. Technical report, Biowulf Technologies, 2001.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag, New York, 1984.

B. E. Boser, I. M. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. D. Jackel, Y. LeCun, U. A. Müller, E. Säckinger, P. Simard, and V. Vapnik. Comparison of classifier methods: a case study in handwritten digit recognition. In *Proceedings of the 12th International Conference on Pattern Recognition and Neural Networks, Jerusalem*, pages 77–87. IEEE Computer Society Press, 1994.

O. Bousquet. PhD thesis, Ecole Polytechnique, 2002.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2001.

C. J. C. Burges and B. Schölkopf. Improving the accuracy and speed of support vector learning machines. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 375–381, Cambridge, MA, 1997. MIT Press.

O. Chapelle and B. Schölkopf. Incorporating invariances in nonlinear SVMs. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002. To appear.

N. Cristianini, A. Elisseeff, and J. Shawe-Taylor. On optimizing kernel alignment. Technical Report 2001-087, NeuroCOLT, 2001.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, UK, 2000.

D. DeCoste and B. Schölkopf. Training invariant support vector machines. *Machine Learning*, 46:161–190, 2002. Also: Technical Report JPL-MLTR-00-1, Jet Propulsion Laboratory, Pasadena, CA, 2000.

F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

L. Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. In M. Li and A. Maruoka, editors, *Algorithmic Learning Theory ALT-97*, LNAI-1316, pages 352–363, Berlin, 1997. Springer.

S. Harmeling, A. Ziehe, M. Kawanabe, and K.-R. Müller. Kernel feature spaces and nonlinear blind source separation. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002. To appear.

D. Haussler. Convolutional kernels on discrete structures. Technical Report UCSC-CRL-99-10, Computer Science Department, University of California at Santa Cruz, 1999.

P. Hayton, B. Schölkopf, L. Tarassenko, and P. Anuzis. Support vector novelty detection applied to jet engine vibration spectra. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 946–952. MIT Press, 2001.

T. S. Jaakkola and D. Haussler. Probabilistic kernel regression models. In *Proceedings of the 1999 Conference on AI and Statistics*, 1999.

G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41:495–502, 1970.

V. Kolchinskiĭ, D. Panchenko, and F. Lozano. Further explanation of the effectiveness of voting methods: The game between margins and weights. In D. Helmbold and R. C. Williamson, editors, *Proceedings of the 14th Annual Conference on Computational Learning Theory, Amsterdam*, LNCS. Springer, 2001.

I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of ICML'2002*, 2002.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.

N. Littlestone and M. Warmuth. Relating data compression and learnability. Technical report, University of California Santa Cruz, 1986.

H. Lodhi, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. Technical Report 2000-79, NeuroCOLT, 2000. Published in: T. K. Leen, T. G. Dietterich and V. Tresp (eds.), *Advances in Neural Information Processing Systems 13*, MIT Press, 2001, as well as in JMLR **2**:419-444, 2002.

D. J. C. MacKay. Introduction to Gaussian processes. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, pages 133–165. Springer-Verlag, Berlin, 1998.

S. Mendelson. Rademacher averages and phase transitions in Glivenko-Cantelli classes. *IEEE Transactions on Information Theory*, 2001. Submitted.

C. A. Micchelli. Algebraic aspects of interpolation. *Proceedings of Symposia in Applied Mathematics*, 36:81–102, 1986.

T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9), 1990.

B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

B. Schölkopf, J. Weston, E. Eskin, C. Leslie, and W. S. Noble. A kernel approach for learning from almost orthogonal patterns. In *Proceedings of the 13th European Conference on Machine Learning (ECML'2002) and Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'2002), Helsinki*, volume 2430/2431 of *Lecture Notes in Computer Science*, Berlin, 2002. Springer.

M. Seeger. Bayesian methods for support vector machines and Gaussian processes. Master's thesis, University of Edinburgh, Division of Informatics, 1999.

J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

P. Simard, Y. LeCun, and J. Denker. Efficient pattern recognition using a new transformation distance. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems 5. Proceedings of the 1992 Conference*, pages 50–58, San Mateo, CA, 1993. Morgan Kaufmann.

A. J. Smola and B. Schölkopf. On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 22:211–231, 1998.

K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg, and K.R. Müller. A new discriminative kernel from probabilistic models. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.

V. Vapnik. *Estimation of Dependences Based on Empirical Data [in Russian]*. Nauka, Moscow, 1979. (English translation: Springer Verlag, New York, 1982).

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, Moscow, 1974. (German Translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie–Verlag, Berlin, 1979).

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

J.-P. Vert. A tree kernel to analyze phylogenetic profiles. In *Proceedings of ISMB'02*, 2002.

U. von Luxburg, O. Bousquet, and B. Schölkopf. A compression approach to support vector model selection. Technical report, Max Planck Institute for Biological Cybernetics, 2002.

G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1990.

C. Watkins. Dynamic alignment kernels. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 39–50, Cambridge, MA, 2000. MIT Press.

H. L. Weinert, editor. *Reproducing Kernel Hilbert Spaces — Applications in Statistical Signal Processing*. Hutchinson Ross, Stroudsburg, PA, 1982.

J. Weston, O. Chapelle, A. Elisseeff, B. Schölkopf, and V. Vapnik. Kernel dependency estimation. Technical Report 98, Max Planck Institute for Biological Cybernetics, 2002.

C. K. I. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In M. I. Jordan, editor, *Learning and Inference in Graphical Models*. Kluwer, 1998.

R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. Technical Report 19, NeuroCOLT, http://www.neurocolt.com, 1998. Accepted for publication in IEEE Transactions on Information Theory.

A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 16(9):799–807, 2000.