Kernel Methods

Martin Sewell
Department of Computer Science
University College London

April 2007 (revised July 2009)

1 Terminology

The term *kernel* is derived from a word that can be traced back to *c.* 1000 and originally meant a seed (contained within a fruit) or the softer (usually edible) part contained within the hard shell of a nut or stone-fruit. The former meaning is now obsolete. It was first used in mathematics when it was defined for integral equations in which the kernel is known and the other function(s) unknown, but now has several meanings in maths. The machine learning term *kernel trick* was first used in 1998.

2 Definition

The *kernel* of a function f is the equivalence relation on the function's domain that roughly expresses the idea of 'equivalent as far as the function f can tell'.

Definition 1 *Let X and Y be sets and let f be a function from X to Y. Elements x_1 and x_2 of X are equivalent if $f(x_1)$ and $f(x_2)$ are equal, i.e. are the same element of Y. Formally: $f : X \to Y$*

$$ker(f) = \{(x_1, x_2) \in X \times X : f(x_1) = f(x_2)\}.$$

The kernel trick (described in Section 5 (page 6)) uses the kernel as a similarity measure and the term *kernel function* is often used for f above.

3 Motivation and Description

Firstly, linearity is rather special, and outside quantum mechanics no model of a real system is truly linear. Secondly, detecting linear relations has been the focus of much research in statistics and machine learning for decades and the resulting algorithms are well understood, well developed and efficient. Naturally, one wants the best of both worlds. So, if a problem is non-linear, instead of trying to fit a non-linear model, one can map the problem from the *input space*

to a new (higher-dimensional) space (called the *feature space*) by doing a non-linear transformation using suitably chosen basis functions and then use a linear model in the feature space. This is known as the 'kernel trick'. The linear model in the feature space corresponds to a non-linear model in the input space. This approach can be used in both classification and regression problems. The choice of kernel function is crucial for the success of all kernel algorithms because the kernel constitutes prior knowledge that is available about a task. Accordingly, there is no free lunch in kernel choice.

4 History

David Hilbert used the German word *kern* in his first paper on integral equations (Hilbert 1904).

The mathematical result underlying the kernel trick, Mercer's theorem, is almost a century old (Mercer 1909). It tells us that any 'reasonable' kernel function corresponds to *some* feature space.

The underlying mathematical results that allow us to determine which kernels can be used to compute distances in feature spaces was developed by Schoenberg (1938).

The methods for representing kernels in linear spaces were first studied by Kolmogorov (1941) for a countable input domain.

The method for representing kernels in linear spaces for the general case was developed by Aronszajn (1950).

Dunford and Schwartz (1963) showed that Mercer's theorem also holds true for general compact spaces.

The use of Mercer's theorem for interpreting kernels as inner products in a feature space was introduced into machine learning by Aizerman, Braverman and Rozonoer (1964).

The justification for a non-linear transformation followed by a linear transformation can be traced back to Cover (1965).

The idea of polynomial kernels stems from Poggio (1975).

Berg, Christensen and Ressel (1984) published a good monograph on the theory of kernels.

Micchelli (1986) discussed closure properties when making kernels.

Saitoh (1988) showed the connection between positivity (a 'positive matrix' defined in Aronszajn (1950)) and the positive semi-definiteness of all finite set kernel matrices.

Reproducing kernels were extensively used in machine learning and neural networks by Poggio and Girosi, see for example Poggio and Girosi (1990), a paper on radial basis function networks. The theory of kernels was used in approximation and regularisation theory, and the first chapter of *Spline Models for Observational Data* (Wahba 1990) gives a number of theoretical results on kernel functions.

In a seminal paper, Boser, Guyon and Vapnik (1992) (re)introduced the notion of a kernel into the mainstream of the machine learning literature by

combining kernel functions with large margin hyperplanes, leading to support vector machines. They discussed Gaussian and polynomial kernels.

ANOVA kernels were first suggested by Burges and Vapnik (1995) (under the name *Gabor kernels*). FitzGerald, Micchelli and Pinkus (1995) studied various notions of multivariate functions which map families of positive semidefinite matrices or of conditionally positive semidefinite matrices into matrices of the same type.

Schölkopf, Smola and Müller (1996) used kernel functions to perform principal component analysis.

For a very readable book covering kernels for strings and trees, see Gusfield (1997). Schölkopf (1997) observed that *any* algorithm which can be formulated solely in terms of dot products can be made non-linear by carrying it out in feature spaces induced by Mercer kernels. Schölkopf, Smola and Müller (1997) presented their paper on kernel PCA.

Smola, Schölkopf and Müller (1998) make the connection between regularization operators and support vector kernels. An early survey of the modern usage of kernel methods in pattern analysis can be found in Burges (1998). For another readable book on kernels for strings and trees, see Durbin, *et al.* (1999). Graepel and Obermayer (1998) proposed a topographic clustering algorithm based on kernel functions. Williams (1998) gives a tutorial on regression with Gaussian processes. Smola and Schölkopf (1998) generalized the support vector approach to a wider range of cost functions, and established a link between regularization operators and support vector kernels. Vapnik (1998) described *recursive ANOVA kernels*. Joachims (1998) proposed *bag-of-word kernels*, which can be considered as an example of kernels between sets. MacKay (1998) gives an introduction to Gaussian processes. Schölkopf, Burges and Smola (1998) edited *Advances in Kernel Methods: Support Vector Learning*, which is a collection of papers submitted during a workshop on SVMs held at the 1997 annual Neural Information Processing Systems (NIPS) conference. Schölkopf, Smola and Müller (1998) published their work on kernel principal component analysis. Stitson, *et al.* (1998) used ANOVA decomposition kernels to good effect in a multi-dimensional support vector regression problem. Wahba (1998) gives a survey of work on reproducing kernel Hilbert spaces.

Watkins (1999b) proposed the use of probabilistic context-free grammars to build kernels between sequences. Evgeniou, Pontil and Poggio (1999) present a unified framework for regularization networks and SVMs. Cristianini, Campbell and Shawe-Taylor (1999) presented an algorithm which automatically learns the kernel parameter from the data. Jaakkola and Haussler (1999a) proposed using a hidden Markov model to evaluate a kernel between biosequences, where the feature vector is the Fisher score of the distribution; they introduced the *Fisher kernel*. Jaakkola and Haussler (1999b) derived a generic class of probabilistic regression models and a parameter estimation technique that can make use of arbitrary kernel functions. Amari and Wu (1999) described a method for modifying a kernel function in order to affect the geometry in the input space, so that the separability of the data is improved. Haussler (1999) proposed a new method of constructing kernels on sets whose elements are discrete struc-

tures like strings, trees and graphs. He introduced P-kernels. Watkins (1999a) developed a string subsequence kernel by means of recursion.

Cristianini and Shawe-Taylor (2000) published *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods* which includes a chapter on Kernel-Induced Feature Spaces. Hofmann (2000) used the Fisher kernel for learning the similarity between text documents. Watkins (2000) showed that the scores produced by certain dynamic alignment algorithms for sequences are in fact valid kernel functions. Zien, *et al.* (2000) showed how to incorporate prior biological knowledge by engineering an appropriate kernel function, and developed a *locality-improved kernel*. Oliver, Schölkopf and Smola (2000) provided a regularization-theoretic analysis of a class of SV kernels—called *natural kernels*—based on generative models with density $p(\mathbf{x}|\theta)$, such as the Fisher kernel.

Sim (2001) developed a kernel for text categorization which looks only at pairs of words within a certain vicinity with respect to each other. Bartlett and Schölkopf (2001) consider some kernels for structured data. Williamson, Smola and Schölkopf (2001) derive new bounds for the generalization error of kernel machines. Steinwart (2001) considered the influence of the kernel on the consistency of SVMs and developed the concept of *universal kernels*. Herbrich (2001) published the book *Learning Kernel Classifiers: Theory and Algorithms* which provided the first comprehensive overview of both the theory and algorithms of kernel classifiers. Schölkopf and Smola (2001) wrote the book *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, which provides an introduction to SVMs and related kernel methods, and may appeal to mathematicians.

Lodhi, *et al.* (2002) were the first to apply string kernels, with an application to text categorisation. Leslie, Eskin and Noble (2002) introduced a new sequence-similarity (string) kernel, the *spectrum kernel*, for use with SVMs in a discriminative approach to the protein classification problem. Vert (2002b) defined a class of kernels for strings based on simple probabilistic models. In a text classification problem, Joachims (2002) proposed using the vector space model as a kernel. Girolami (2002) applied the kernel trick to clustering in feature space. Gärtner, *et al.* (2002) introduced *multi-instance kernels*, kernels for multi-instance problems—a class of concepts on individuals represented by sets. Kondor and Lafferty (2002) proposed *diffusion kernels*. Tsuda, Kin and Asai (2002) proposed marginalised kernels, which provide a way to design a kernel from latent variables. Vert (2002a) introduced the *tree kernel*. Takimoto and Warmuth (2002) introduced the *all subsets kernel* as well as the idea of kernels based on paths in a graph. In an application of kernel methods to natural language processing problems, Collins and Duffy (2002) presented the dynamic programming method for comparing two trees in the context of language parsing. Steinwart (2002) showed that SVMs of the 1-norm soft margin type are universally consistent provided that the regularization parameter is chosen in a distinct manner and the kernel belongs to the class of universal kernels.

Jordan (2003) wrote an introduction to probabilistic graphical models. The survey paper by Gärtner (2003) describes several approaches to defining positive

definite kernels on structured data. Cortes, Haffner and Mohri (2003b) discuss positive definite rational kernels. Jebara and Kondor (2003) introduced a class of kernels between probability distributions, the *probability product kernel*, which eschews some of the complexities that kernels based on the Kullback-Leibler divergence often contend with. The *Bhattacharyya kernel* and the *expected likelihood kernel* are special cases. Leslie and Kuang (2003) introduced several new families of string kernels—*restricted gappy kernels*, *substitution kernels* and *wildcard kernels*—designed in particular for use with SVMs for classification of protein sequence data. Smola and Kondor (2003) introduced a family of kernels on graphs based on the notion of regularization operators. Cortes, Haffner and Mohri (2003a) introduced a general family of kernels based on weighted transducers or rational relations, *rational kernels*. Borrowing ideas and techniques from information theory and data compression, Cuturi and Vert (2003) proposed a covariance kernel for biological sequences. Kandola, Shawe-Taylor and Cristianini (2003) introduced the *von Neumann kernel*. Leslie, *et al.* (2003) introduced a class of string kernels called *mismatch kernels*. Saunders, Shawe-Taylor and Vinokourov (2003) showed how string kernels can be treated as Fisher kernels. Vishwanathan and Smola (2003) described novel methods of computation using suffix trees. Wolf and Shashua (2003) considered the problem of learning with instances defined over a space of sets of vectors. They derived a positive definite kernel defined over pairs of matrices based on the concept of principal angles between two linear subspaces.

Lanckriet, *et al.* (2004) show how the kernel matrix can be learned from data via semidefinite programming techniques. Kashima, Tsuda and Inokuchi (2004a) introduced a kernel function between two labelled graphs in the framework of marginalized kernels. Leslie, *et al.* (2004) introduced a class of string kernels, called *mismatch kernels*. Pinkus (2004) proved a result concerning conditions in which a function is strictly positive definite. Shawe-Taylor and Cristianini (2004) wrote *Kernel Methods for Pattern Analysis*, which provides practitioners with a large toolkit of algorithms, kernels and solutions ready to be implemented, and also serves as an introduction for students and researchers to kernel-based pattern analysis. The book may appeal to those with a computer science bent. Kashima, Tsuda and Inokuchi (2004b) discussed the construction of kernel functions between labelled graphs and provide a unified account of a family of kernels called *label sequence kernels*. Vert, Saigo and Akutsu (2004) used pair hidden Markov models as kernels. Vishwanathan and Smola (2004) presented a new algorithm suitable for matching discrete objects such as strings and trees in linear time, thus obviating dynamic programming with quadratic time complexity.

Hein, Bousquet and Schölkopf (2005) built a general framework for the generation of maximal margin algorithms for metric spaces. Rasmussen and Williams (2005) provide a systematic and unified treatment of theoretical and practical aspects of Gaussian processes in machine learning.

Borgwardt, *et al.* (2006) proposed a kernel-based statistical test of whether two samples are from the same distribution.

Vishwanathan, Smola and Vidal (2007) proposed a family of kernels based

on the Binet-Cauchy theorem, and its extension to Fredholm operators. Their derivation provides a unifying framework for all kernels on dynamical systems currently used in machine learning, including kernels derived from the behavioural framework, diffusion processes, marginalized kernels, kernels on graphs and the kernels on sets arising from the subspace angle approach. Bakir, *et al.* (2007) edited *Predicting Structured Data*.

Filippone, *et al.* (2008) give a a survey of kernel and spectral methods for clustering, and prove that the two paradigms have the same objective. Hofmann, Schölkopf and Smola (2008) review machine learning methods employing positive definite kernels.

5 Kernel Trick

The kernel trick was first published by Aizerman, Braverman and Rozonoer (1964). Mercer's theorem states that any continuous, symmetric, positive semi-definite kernel function $K(x, y)$ can be expressed as a dot product in a high-dimensional space.

If the arguments to the kernel are in a measurable space X, and if the kernel is positive semi-definite—i.e.

$$\sum_{i,j} K(x_i, x_j) c_i c_j \geq 0$$

for any finite subset $\{x_1, \ldots, x_n\}$ of X and subset $\{c_1, \ldots, c_n\}$ of objects (typically real numbers or even molecules)—then there exists a function $\varphi(x)$ whose range is in an inner product space of possibly high dimension, such that

$$K(x, y) = \varphi(x) \cdot \varphi(y).$$

6 Advantages

- The kernel defines a similarity measure between two data points and thus allows one to incorporate prior knowledge of the problem domain.
- Most importantly, the kernel contains all of the information about the relative positions of the inputs in the feature space and the actual learning algorithm is based only on the kernel function and can thus be carried out without explicit use of the feature space. The training data only enter the algorithm through their entries in the kernel matrix (a Gram matrix), and never through their individual attributes. Because one never explicitly has to evaluate the feature map in the high dimensional feature space, the kernel function represents a computational shortcut.
- The number of operations required is not necessarily proportional to the number of features.

References

AIZERMAN, M. A., E. M. BRAVERMAN, and L. I. ROZONOER, 1964. Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning. *Automation and Remote Control*, **25**(6), 821–837.

AMARI, S., and S. WU, 1999. Improving Support Vector Machine Classifiers by Modifying Kernel Functions. *Neural Networks*, **12**(6), 783–789.

ARONSZAJN, N., 1950. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, **68**(3), 337–404.

BAKIR, Gökhan H., *et al.*, eds., 2007. *Predicting Structured Data*. Neural Information Processing. Cambridge, MA: The MIT Press.

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

BERG, Christian, Jens Peter Reus CHRISTENSEN, and Paul RESSEL, 1984. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Volume 100 of *Graduate Texts in Mathematics*. New York: Springer-Verlag.

BORGWARDT, Karsten M., *et al.*, 2006. Integrating Structured Biological Data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, **22**(14), e49–e57.

BOSER, Bernhard E., Isabelle M. GUYON, and Vladimir N. VAPNIK, 1992. A Training Algorithm for Optimal Margin Classifiers. *In*: *COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. New York: ACM Press, pp. 144–152.

BURGES, Christopher J. C., 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, **2**(2), 121–167.

BURGES, C. J. C., and V. VAPNIK, 1995. A new method for constructing artificial neural networks. Interim technical report N00014-94-C-0186, AT&T Bell Laboratories, Holmdel, NJ.

COLLINS, Michael, and Nigel DUFFY, 2002. Convolution Kernels for Natural Language. *In*: Thomas G. DIETTERICH, Suzanna BECKER, and Zoubin GHAHRAMANI, eds. *Advances in Neural Information Processing Systems 14*, Bradford Books. Neural Information Processing. Cambridge, MA: The MIT Press, pp. 625–632.

CORTES, Corinna, Patrick HAFFNER, and Mehryar MOHRI, 2003a. Rational Kernels. *In*: Suzanna BECKER, Sebastian THRUN, and Klaus OBERMAYER, eds. *Advances in Neural Information Processing Systems 15*, Bradford Books. Cambridge, MA: The MIT Press, pp. 617–624.

CORTES, Corinna, Patrick HAFFNER, and Mehryar MOHRI, 2003b. Positive Definite Rational Kernels. *In*: Bernhard SCHÖLKOPF and Manfred K. WARMUTH, eds. *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 2003, Proceedings*, Volume 2777 of *Lecture Notes in Computer Science*. Berlin: Springer, pp. 41–56.

COVER, Thomas M., 1965. Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Transactions on Electronic Computers*, **14**(3), 326–334.

CRISTIANINI, Nello, Colin CAMPBELL, and John SHAWE-TAYLOR, 1999. Dynamically Adapting Kernels in Support Vector Machines. *In*: Michael S. KEARNS, Sara A. SOLLA, and David A. COHN, eds. *Advances in Neural Information Processing Systems 11*, Bradford Books. Cambridge, MA: The MIT Press, pp. 204–210.

CRISTIANINI, Nello, and John SHAWE-TAYLOR, 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press.

CUTURI, Marco, and Jean-Philippe VERT, 2003. A Covariance Kernel for Proteins. arXiv:q-bio/0310022.

DUNFORD, Nelson, and Jacob T. SCHWARTZ, 1963. *Linear Operators Part II Spectral Theory: Self Adjoint Operators in Hilbert Space*. Volume VII of *Pure and Applied Mathematics*. New York: Wiley.

DURBIN, R., *et al.*, 1999. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.

EVGENIOU, Theodoros, Massimiliano PONTIL, and Tomaso POGGIO, 1999. A Unified Framework for Regularization Networks and Support Vector Machines. Technical Report A.I. Memo No. 1654/C.B.C.L Paper No. 171, Massachusetts Institute of Technology, Cambridge, MA.

FILIPPONE, Maurizio, *et al.*, 2008. A Survey of Kernel and Spectral Methods for Clustering. *Pattern Recognition*, **41**(1), 176–190.

FITZGERALD, Carl H., Charles A. MICCHELLI, and Allan PINKUS, 1995. Functions That Preserve Families of Positive Semidefinite Matrices. *Linear Algebra and its Applications*, **221**, 83–102.

GÄRTNER, Thomas, 2003. Kernel-Based Learning in Multi-Relational Data Mining. *SIGKDD Explorations*, **5**(1), 49–58.

GÄRTNER, Thomas, *et al.*, 2002. Multi-Instance Kernels. *In*: Claude SAMMUT and Achim G. HOFFMANN, eds. *Proceedings of the Nineteenth International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann, pp. 179–186.

GIROLAMI, M., 2002. Mercer Kernel-Based Clustering in Feature Space. *IEEE Transactions on Neural Networks*, **13**(3), 780–784.

GRAEPEL, T., and K. OBERMAYER, 1998. Fuzzy Topographic Kernel Clustering. *In*: W. BRAUER, ed. *Proceedings of the 5th GI Workshop Fuzzy Neuro Systems '98*. pp. 90–97.

GUSFIELD, Dan, 1997. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge: Cambridge University Press.

HAUSSLER, David, 1999. Convolution Kernels on Discrete Structures. Technical Report UCSC-CRL-99-10, University of California, Santa Cruz.

HEIN, Matthias, Olivier BOUSQUET, and Bernhard SCHÖLKOPF, 2005. Maximal Margin Classification for Metric Spaces. *Journal of Computer and System Sciences*, **71**(3), 333–359.

HERBRICH, Ralf, 2001. *Learning Kernel Classifiers: Theory and Algorithms*. Adaptive Computation and Machine Learning. Cambridge, MA: The MIT Press.

HILBERT, David, 1904. Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen. (Erste Mitteilung.). *Nachrichten von der Königl. Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse*, **1904**(1), 49–91. In German.

HOFMANN, Thomas, 2000. Learning the Similarity of Documents: An Information-Geometric Approach to Document Retrieval and Categorization. *In*: Sara A. SOLLA, Todd K. LEEN, and Klaus-Robert MÜLLER, eds. *Advances in Neural Information Processing Systems 12*. Cambridge, MA: The MIT Press, pp. 914–920.

HOFMANN, Thomas, Bernhard SCHÖLKOPF, and Alexander J. SMOLA, 2008. Kernel Methods in Machine Learning. *The Annals of Statistics*, **36**(3), 1171–1220.

JAAKKOLA, Tommi S., and David HAUSSLER, 1999a. Exploiting Generative Models in Discriminative Classifiers. *In*: Michael S. KEARNS, Sara A. SOLLA, and David A. COHN, eds. *Advances in Neural Information Processing Systems 11*, Bradford Books. Cambridge, MA: The MIT Press, pp. 487–493.

JAAKKOLA, Tommi S., and David HAUSSLER, 1999b. Probabilistic Kernel Regression Models. *In*: David HECKERMAN and Joe WHITTAKER, eds. *Proceedings of the 1999 Conference on AI and Statistics*. San Mateo, CA: Morgan Kaufmann.

JEBARA, Tony, and Risi KONDOR, 2003. Bhattacharyya and Expected Likelihood Kernels. *In*: Bernhard SCHÖLKOPF and Manfred K. WARMUTH, eds. *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 2003, Proceedings*, Volume 2777 of *Lecture Notes in Artificial Intelligence*. Berlin: Springer-Verlag, pp. 57–71.

JOACHIMS, Thorsten, 1998. Text Categorization with Support Vector Machines: Learning with many Relevant Features. *In*: Claire NÉDELLEC and Céline ROUVEIROL, eds. *Machine Learning: ECML-98: 10th European Conference on Machine Learning, Chemnitz, Germany, April 1998. Proceedings*, Volume 1398 of *Lecture Notes in Computer Science*. Berlin: Springer-Verlag, pp. 137–142.

JOACHIMS, Thorsten, 2002. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. The Kluwer International Series in Engineering and Computer Science. Boston, MA: Kluwer Academic Publishers.

JORDAN, Michael I., 2003. An Introduction to Probabilistic Graphical Models. Unpublished manuscript.

KANDOLA, Jaz, John SHAWE-TAYLOR, and Nello CRISTIANINI, 2003. Learning Semantic Similarity. *In*: Suzanna BECKER, Sebastian THRUN, and Klaus OBERMAYER, eds. *Advances in Neural Information Processing Systems 15*, Bradford Books. Cambridge, MA: The MIT Press, pp. 657–664.

KASHIMA, Hisashi, Koji TSUDA, and Akihiro INOKUCHI, 2004a. Marginalized Kernels Between Labeled Graphs. *In*: Tom FAWCETT and Nina MISHRA, eds. *Proceedings of the Twentieth International Conference on Machine Learning*. Menlo Park, CA: AAAI Press, pp. 321–328.

KASHIMA, Hisashi, Koji TSUDA, and Akihiro INOKUCHI, 2004b. Kernels for Graphs. *In*: Bernhard SCHÖLKOPF, Koji TSUDA, and Jean-Philippe VERT, eds. *Kernel Methods in Computational Biology*, Bradford Books. Computational Molecular Biology. Cambridge, MA: The MIT Press, pp. 155–170.

KOLMOGOROV, A. N., 1941. Stationary sequences in Hilbert spaces. *Moscow University Mathematics Bulletin*, **2**(6), 1–40. In Russian.

KONDOR, Risi Imre, and John LAFFERTY, 2002. Diffusion Kernels on Graphs and Other Discrete Structures. *In*: Andrea DANYLUK, ed. *ICML '02: 19th International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann, pp. 315–322.

LANCKRIET, Gert R. G., *et al.*, 2004. Learning the Kernel Matrix with Semidefinite Programming. *Journal of Machine Learning Research*, **5**, 27–72.

LESLIE, Christina, Eleazar ESKIN, and William Stafford NOBLE, 2002. The Spectrum Kernel: A String Kernel for SVM Protein Classification. *In*: Russ B. ALTMAN, *et al.*, eds. *Pacific Symposium on Biocomputing 2002*. Singapore: World Scientific, pp. 564–575.

LESLIE, Christina, *et al.*, 2003. Mismatch String Kernels for SVM Protein Classification. *In*: Suzanna BECKER, Sebastian THRUN, and Klaus OBERMAYER, eds. *Advances in Neural Information Processing Systems 15*, Bradford Books. Cambridge, MA: The MIT Press, pp. 1441–1448.

LESLIE, Christina, and Rui KUANG, 2003. Fast Kernels for Inexact String Matching. *In*: Bernhard SCHÖLKOPF and Manfred K. WARMUTH, eds. *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 2003, Proceedings*, Volume 2777 of *Lecture Notes in Computer Science*. Berlin: Springer-Verlag, pp. 114–128.

LESLIE, Christina S., *et al.*, 2004. Mismatch String Kernels for Discriminative Protein Classification. *Bioinformatics*, **20**(4), 467–476.

LODHI, Huma, *et al.*, 2002. Text Classification using String Kernels. *Journal of Machine Learning Research*, **2**, 419–444.

MACKAY, David J. C., 1998. Introduction to Gaussian Processes. *In*: Christopher M. BISHOP, ed. *Neural Networks and Machine Learning*, Volume 168 of *NATO ASI Series. Series F, Computer and Systems Sciences*. Berlin: Springer, pp. 133–165.

MERCER, J., 1909. Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations. *Philosophical Transactions of the Royal Society of London. Series A. Containing Papers of a Mathematical or Physical Character*, **209**, 415–446.

MICCHELLI, Charles A., 1986. Interpolation of Scattered Data: Distance Matrices and Conditionally Positive Definite Functions. *Constructive Approximation*, **2**(1), 11–22.

OLIVER, Nuria, Bernhard SCHÖLKOPF, and Alexander J. SMOLA, 2000. Natural Regularization from Generative Models. *In*: Alexander J. SMOLA, *et al.*, eds. *Advances in Large Margin Classifiers*, Neural Information Processing. Cambridge, MA: The MIT Press, Chapter 4, pp. 51–60.

PINKUS, Allan, 2004. Strictly Positive Definite Functions on a Real Inner Product Space. *Advances in Computational Mathematics*, **20**(4), 263–271.

POGGIO, T., 1975. On Optimal Nonlinear Associative Recall. *Biological Cybernetics*, **19**(4), 201–209.

POGGIO, Tomaso, and Federico GIROSI, 1990. Networks for Approximation and Learning. *Proceedings of the IEEE*, **78**(9), 1481–1497.

RASMUSSEN, Carl Edward, and Christopher K. I. WILLIAMS, 2005. *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press.

SAITOH, Saburou, 1988. *Theory of Reproducing Kernels and its Applications*. Pitman Research Notes in Mathematics. Harlow: Longman Scientific & Technical.

SAUNDERS, Craig, John SHAWE-TAYLOR, and Alexei VINOKOUROV, 2003. String Kernels, Fisher Kernels and Finite State Automata. *In*: Suzanna BECKER, Sebastian THRUN, and Klaus OBERMAYER, eds. *Advances in Neural Information Processing Systems 15*, Bradford Books. Cambridge, MA: The MIT Press, pp. 633–640.

SCHOENBERG, I. J., 1938. Metric Spaces and Positive Definite Functions. *Transactions of the American Mathematical Society*, **44**(3), 522–536.

SCHÖLKOPF, Bernhard, 1997. *Support Vector Learning*. Ph. D. thesis, Technische Universität Berlin, Berlin. Published by R. Oldenbourg Verlag, Munich.

SCHÖLKOPF, Bernhard, Christopher J. C. BURGES, and Alexander J. SMOLA, eds., 1998. *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA: The MIT Press.

SCHÖLKOPF, Bernhard, Alexander SMOLA, and Klaus-Robert MÜLLER, 1996. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Technical Report 44, Max-Planck-Institut für biologische Kybernetik Arbeitsgruppe Bülthoff, Tübingen.

SCHÖLKOPF, Bernhard, Alexander SMOLA, and Klaus-Robert MÜLLER, 1997. Kernel Principal Component Analysis. *In*: Wulfram GERSTNER, *et al.*, eds. *Artificial Neural Networks – ICANN '97: 7th International Conference, Lausanne, Switzerland, October 1997, Proceedings*, Volume 1327 of *Lecture Notes in Computer Science*. Berlin: Springer, pp. 583–588.

SCHÖLKOPF, Bernhard, and Alexander J. SMOLA, 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. Cambridge, MA: The MIT Press.

SCHÖLKOPF, Bernhard, Alexander J. SMOLA, and Klaus-Robert MÜLLER, 1998. Kernel Principal Component Analysis. *In*: Bernhard SCHÖLKOPF, Christopher J. C. BURGES, and Alexander J. SMOLA, eds. *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA: The MIT Press, Chapter 20, pp. 327–352. Short version published in 1998 as Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation* **10**(5), 1299–1319.

SHAWE-TAYLOR, John, and Nello CRISTIANINI, 2004. *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press.

SIM, Kristy, 2001. Context kernels for text categorization. Master's thesis, The Australian National University, Canberra.

SMOLA, Alexander J., and Risi KONDOR, 2003. Kernels and Regularization on Graphs. *In*: Bernhard SCHÖLKOPF and Manfred K. WARMUTH, eds. *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 2003, Proceedings*, Volume 2777 of *Lecture Notes in Artificial Intelligence*. Berlin: Springer-Verlag, pp. 144–158.

SMOLA, A. J., and B. SCHÖLKOPF, 1998. On a Kernel-Based Method for Pattern Recognition, Regression, Approximation, and Operator Inversion. *Algorithmica*, **22**(1–2), 211–231. First released as GMD Technical Report No. 1064 (1997).

SMOLA, Alex J., Bernhard SCHÖLKOPF, and Klaus-Robert MÜLLER, 1998. The Connection Between Regularization Operators and Support Vector Kernels. *Neural Networks*, **11**(4), 637–649.

STEINWART, Ingo, 2001. On the Influence of the Kernel on the Consistency of Support Vector Machines. *Journal of Machine Learning Research*, **2**, 67–93.

STEINWART, Ingo, 2002. Support Vector Machines are Universally Consistent. *Journal of Complexity*, **18**(3), 768–791.

STITSON, Mark O., *et al.*, 1998. Support Vector Regression with ANOVA Decomposition Kernels. *In*: Bernhard SCHÖLKOPF, Christopher J. C. BURGES, and Alexander J. SMOLA, eds. *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA: The MIT Press, Chapter 17, pp. 285–291.

TAKIMOTO, Eiji, and Manfred K. WARMUTH, 2002. Path Kernels and Multiplicative Updates. *In*: Jyrki KIVINEN and Robert H. SLOAN, eds. *Computational Learning Theory: 15th Annual Conference on Computational Learning Theory, COLT 2002, Sydney, Australia, July 2002, Proceedings*, Volume 2375 of *Lecture Notes in Computer Science*. Berlin: Springer-Verlag, pp. 74–89.

TSUDA, Koji, Taishin KIN, and Kiyoshi ASAI, 2002. Marginalized Kernels for Biological Sequences. *Bioinformatics*, **18**(Supplement 1), S268–S275.

VAPNIK, Vladimir N., 1998. *Statistical Learning Theory*. New York: Wiley.

VERT, Jean-Philippe, 2002a. A Tree Kernel to Analyse Phylogenetic Profiles. *Bioinformatics*, **18**(Supplement 1), S276–S284.

VERT, J.-P., 2002b. Support Vector Machine Prediction of Signal Peptide Cleavage Site Using a New Class of Kernels for Strings. *In*: Russ B. ALTMAN, *et al.*, eds. *Pacific Symposium on Biocomputing 2002*. Singapore: World Scientific Publishing, pp. 649–660.

VERT, Jean-Philippe, Hiroto SAIGO, and Tatsuya AKUTSU, 2004. Local alignment kernels for biological sequences. *In*: Bernhard SCHÖLKOPF, Koji TSUDA, and Jean-Philippe VERT, eds. *Kernel Methods in Computational Biology*, Bradford Books. Computational Molecular Biology. Cambridge, MA: The MIT Press, Chapter 6, pp. 131–154.

VISHWANATHAN, S. V. N., and Alexander J. SMOLA, 2003. Fast Kernels for String and Tree Matching. *In*: Suzanna BECKER, Sebastian THRUN, and Klaus OBERMAYER, eds. *Advances in Neural Information Processing Systems 15*, Bradford Books. Cambridge, MA: The MIT Press, pp. 585–592.

VISHWANATHAN, S. V. N., and Alexander Johannes SMOLA, 2004. Fast Kernels for String and Tree Matching. *In*: Bernhard SCHÖLKOPF, Koji TSUDA, and Jean-Philippe VERT, eds. *Kernel Methods in Computational Biology*, Bradford Books. Computational Molecular Biology. Cambridge, MA: The MIT Press, Chapter 5, pp. 113–130.

VISHWANATHAN, S. V. N., Alexander J. SMOLA, and René VIDAL, 2007. Binet-Cauchy Kernels on Dynamical Systems and its Application to the Analysis of Dynamic Scenes. *International Journal of Computer Vision*, **73**(1), 95–119.

WAHBA, Grace, 1990. *Spline Models for Observational Data*. Volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Philadelphia, PA: SIAM: Society for Industrial and Applied Mathematics.

WAHBA, Grace, 1998. Support Vector Machines, Reproducing Kernel Hilbert Spaces, and Randomized GACV. *In*: Bernhard SCHÖLKOPF, Christopher J. C. BURGES, and Alexander J. SMOLA, eds. *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA: The MIT Press, Chapter 6, pp. 69–88.

WATKINS, Chris, 1999a. Kernels from Matching Operations. Technical Report CSD-TR-98-07, Royal Holloway, University of London, Egham.

WATKINS, Chris, 1999b. Dynamic Alignment Kernels. Technical Report CSD-TR-98-11, Royal Holloway, University of London, Egham.

WATKINS, Chris, 2000. Dynamic alignment kernels. *In*: Alexander J. SMOLA, *et al.*, eds. *Advances in Large Margin Classifiers*, Neural Information Processing. Cambridge, MA: The MIT Press, Chapter 3, pp. 39–50.

WILLIAMS, C. K. I., 1998. Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond. *In*: Michael I. JORDAN, ed. *Learning in Graphical Models*, Volume 89 of *NATO Science Series D*. Dordrecht: Kluwer Academic Publishers, pp. 599–621. Also published in 1999 by The MIT Press, Cambridge, MA.

WILLIAMSON, Robert C., Alex J. SMOLA, and Bernhard SCHÖLKOPF, 2001. Generalization Performance of Regularization Networks and Support Vector Machines Via Entropy Numbers of Compact Operators. *IEEE Transactions on Information Theory*, **47**(6), 2516–2532.

WOLF, Lior, and Amnon SHASHUA, 2003. Learning over Sets using Kernel Principal Angles. *Journal of Machine Learning Research*, **4**, 913–931.

ZIEN, A., *et al.*, 2000. Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites. *Bioinformatics*, **16**(9), 799–807.