Introduction to Machine Learning

Johan A.K. Suykens

KU Leuven, ESAT-SCD/SISTA
Kasteelpark Arenberg 10
B-3001 Leuven (Heverlee), Belgium
Email: johan.suykens@esat.kuleuven.be

April 2013

1 Scope and context

Classically, within the signal processing community, linear parametric models have been a method of first choice in several applications. Historically, many computationally efficient algorithms have been developed for on-line and adaptive signal processing with e.g. LMS, recursive least squares and Kalman filtering type algorithms [35]. However, more recently considerable progress has been made also on the use of flexible nonlinear models, e.g. related to kernel methods, support vector machines [29, 31, 42, 53, 54, 56, 55, 62, 63] and probabilistic models [20, 37, 41, 43, 44, 51], and the importance of regularization techniques has been realized both in the context of parametric models and non-parametric models. This is witnessed also by the progress in the area of compressed sensing and sparse models [16, 23, 24, 25, 27, 32]. Moreover, many emerging applications in e.g. big data, networks applications, bioinformatics, brain-machine interfaces, are posing new challenges for predictive models towards handling large amounts of data in high dimensional input spaces. In this Machine Learning Section we therefore take a broad view on the subject of signal

processing & machine learning in connection also to other related areas as pattern recognition and neural networks [19, 33, 58], mathematics and statistics [30, 31, 34], optimization [21, 17] and information theory [50] (Figure 1).

In general one distinguishes between different types of learning models, such as supervised, unsupervised and semi-supervised learning [26, 18, 62], various tasks such as e.g. classification, regression, clustering and different types of models, including e.g. linear and nonlinear parametric models, kernel-based models and probabilistic models (Figure 2). For many of the successful methods it is interesting to trace back to the original roots. For on-line learning of linear models in classification problems, the perceptron has originally served as a paradigm. However, soon one has encountered its limitations. In the neural networks area this led to introducing one or more hidden layers with multilayer perceptron neural networks. Backpropagation as the original learning algorithm for such feedforward networks, in its on-line learning form, could be interpreted as an extension of the LMS algorithm as used in adaptive signal processing [64]. On a different track, the perceptron has also been studied within the context of statistical learning theory [60, 61, 62]. Here one is interested in characterizing the generalization error of the model, which is typically expressed in terms of the error on the training data and a complexity term.

Multilayer perceptrons are universal approximators [36] which make them powerful tools to parameterize nonlinear functions. In order to overcome the problem of overfitting with flexible nonlinear models, an important technique to use is regularization [19, 44]. In the objective function one not only minimizes then the error on the training data but one also keeps the estimated parameter values small. This leads to the notion of effective number of parameters which is relevant then to characterize the model complexity, instead of the number of parameters. The flexibility of the model is controlled by the regularization term. In a Bayesian inference and probabilistic modelling picture the regularization term corresponds then to the prior distribution on the unknown parameters. Classical regularization schemes minimize the ℓ_2 norm on the unknown parameters, which is known as ridge regression in statistics and dates back also to ill-posed problems and Tikhonov regularization [59]. In recent years there has been considerable interest in alternative regularization schemes based on ℓ_0, ℓ_1 and ℓ_p regularization to achieve sparsity in the solution vector (Figure 3), in connection also to compressed sensing [16, 23, 24, 25, 32].

Regularization also plays an important role in non-parametric and kernel-based models. The use of positive definite kernels and reproducing kernels dates back to the early work of Mercer, Moore, Aronszajn [45, 46, 14] and are key ingredients within methods of function estimation in reproducing kernel Hilbert spaces, the theory of splines and radial basis function networks [49, 63]. Early use of reproducing kernel Hilbert spaces in signal processing is e.g. [39, 40, 48]. In Gaussian processes the kernel function relates to the correlation function [43, 51]. An increasing and renewed interest in kernel-based methods appeared with the introduction of nonlinear support vector machines for classification and regression [62]. The use of a positive definite kernel is viewed here in connection to a feature map (often called the kernel trick, which relates to the Mercer theorem), where in the primal a constrained optimization problem formulation is given on the model that is expressed in terms of the feature map. The Lagrange dual problem results then into a kernel-based model representation. In standard support vector machines a sparse kernel-based model is then achieved through the choice of the loss function, typically the hinge loss in classification and the epsilon-insensitive loss function in function estimation.

The kernel trick on its own has also been frequently employed to obtain nonlinear kernel versions of classically known linear estimation schemes, e.g. kernel principal component analysis [52] as an extension to the classical linear principal component analysis [38]. Special kernel functions have also been designed to handle specific data types or in specific applications area such as e.g. textmining or bioinformatics [15, 54, 53]. It is also possible to relate kernel functions to probabilistic graphical models and graphs. In least squares support vector machines one works with simple core models within the primal-dual setting for a wide range of problems in supervised and unsupervised learning and beyond [56, 57]. The primal representation relates then to parametric picture, while the dual representation to a non-parametric. Depending on the nature of the given problem (large number of data versus dimensionality of the input space) this choice in representation can be exploited for developing efficient large scale algorithms [56, 57].

An advantage of support vector machines for classification and regression is that the problem is recasted as a convex optimization problem, up to a small amount of tuning parameters of regularization constant(s) and kernel parameter(s). This has been viewed as a considerable advantage over other nonlinear models such as multilayer perceptrons which

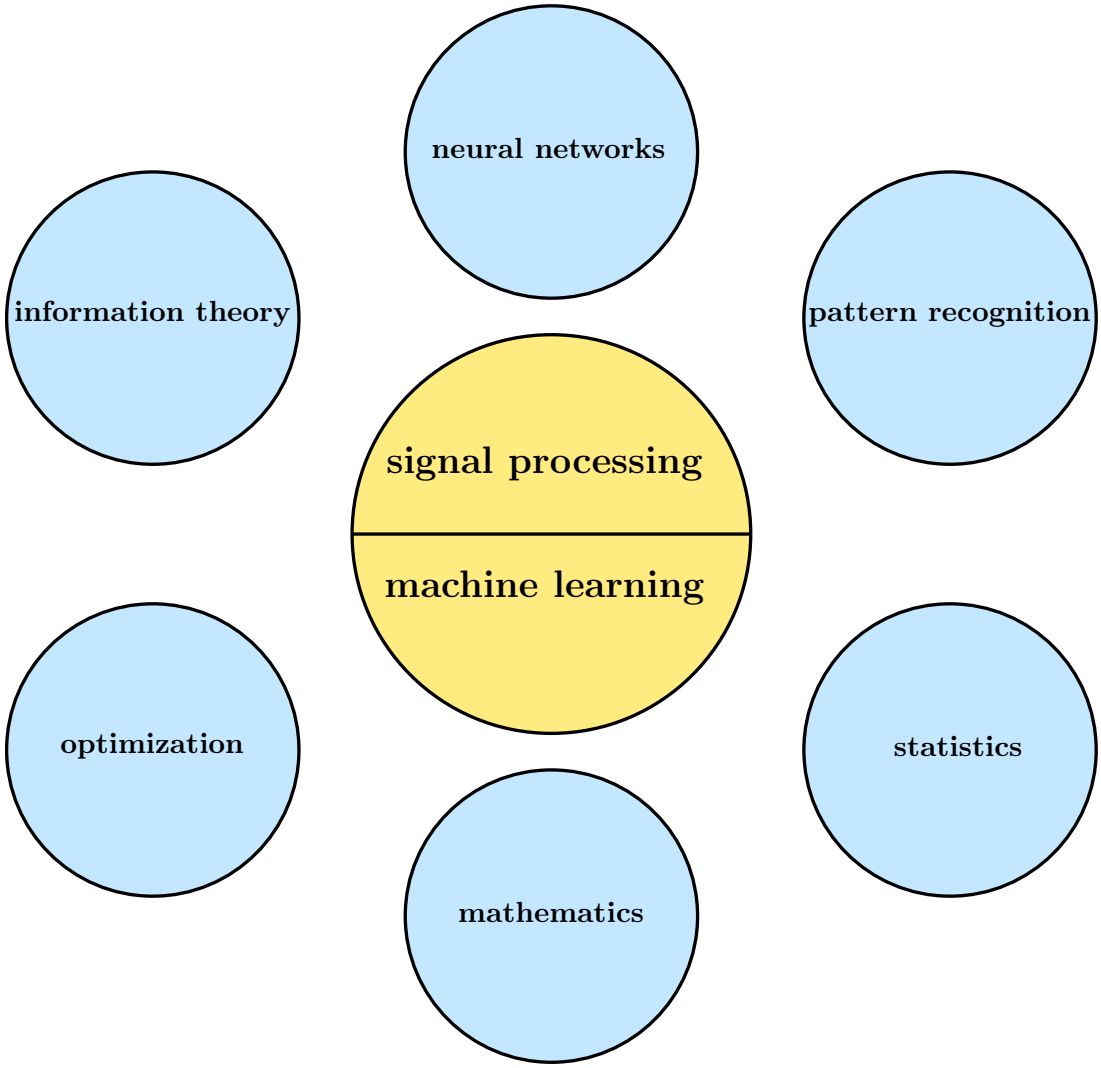

Figure 1: Signal processing & machine learning and several related areas.

suffer from the existence of many local minima solutions. Also towards sparse models and compressed sensing, convex optimization is playing an important role [21] (Figure 3). In many emerging applications one often has to cope with large amounts of data in often high-dimensional input spaces. This is posing new challenges for scalable optimization algorithms. In this direction efficient first order methods, on-line optimization, stochastic optimization or distributed optimization are suitable possible algorithms [22, 47].

In the next Section a brief overview is given on the chapter contributions that present introductory and tutorial contributions related to Signal Processing & Machine Learning.

Learning modes	**Tasks**	**Models**
supervised learning	regression	linear parametric
unsupervised learning	classification	non-linear parametric
semi-supervised learning	clustering	polynomial model
reinforcement learning	density estimation	multilayer perceptron
inductive learning	component analysis	radial basis function network
transductive learning	dimensionality reduction	splines
ensemble learning	data visualization	kernel-based model
transfer learning	manifold learning	support vector machines
	structure/feature selection	graphical models
	multi-task learning	probabilistic models
	dynamical systems modelling	mixture models
	time-series analysis	

Figure 2: Learning modes, learning tasks and examples of different possible models.

Regularization	
Parametric	**Kernel-based**
ℓ_2, ridge regression	RKHS function estimation
ℓ_1, LASSO	splines
ℓ_p $(0 < p \leq 1)$	regularization networks
group LASSO	Gaussian processes
elastic net	support vector machines
spectral regularization	LS-SVMs
nuclear norm	

Figure 3: Regularization and its role in parametric and non-parametric modelling approaches.

2 Contributions

In [1] the authors present an overview of learning theory including statistical and computational aspects, with emphasis on classification and regression problems. Empirical risk minimization is discussed and concepts for characterizing the generalization performance of the model such as Rademacher complexity, covering numbers, Vapnik-Chervonenkis and fat shattering dimension. In connection to this, the problem of model selection is addressed.

In [2] an overview is presented on different types of neural networks for supervised and unsupervised learning. Starting from the perceptron, feedforward networks and backpropagation is explained. Next recurrent neural networks and recursive structure processing are discussed. Neural architectures for principal component analysis and topographic mapping for data mining and data visualization is outlined.

In [3] the authors give an introduction to the foundations and implementations of kernel methods, computational issues and recent developments. This includes the kernel trick, properties and types of kernels, kernel principal component analysis, kernel canonical correlation analysis, kernel Fisher discriminant analysis, support vector machines for classification and regression, and Gaussian processes.

In [4] on-line learning in reproducing kernel Hilbert spaces is presented. First parameter estimation is discussed in regression and classification tasks and how to overcome overfitting by applying regularization. It is explained how a nonlinear task can be mapped to a linear task. In this way kernel LMS and complex kernel LMS are extended to kernel versions of the well-known LMS algorithm in signal processing. For least squares learning algorithms extensions to kernel recursive least squares are discussed. Finally, convex analysis concepts for online learning are provided.

In [5] an introduction to probabilistic graphical models is given. It includes three representations of probabilistic graphical models: Markov networks (or undirected graphical models), Bayesian networks (or directed graphical models) and factor graphs. An overview about structure and parameter learning techniques is given on maximum likelihood and Bayesian learning, and generative and discriminative learning. Exact inference methods and approximate inference techniques are addressed. Applications for each of the three representations are given: Bayesian networks for expert systems, dynamic Bayesian networks for speech processing, Markov random fields for image processing, and factor graphs

for decoding error-correcting codes.

In [6] a tutorial introduction to Monte Carlo Methods, Markov Chain Monte Carlo and Particle Filtering is given. Starting from the Monte Carlo principle and basic techniques for simulating and transforming random variables, Markov Chain Monte Carlo is explained. Other topics that are addressed are rejection sampling, detailed balance, the Gibbs sampler, sequential Monte Carlo, importance sampling, resampling and advanced Monte Carlo methods.

In [7] an introduction to clustering is given. Different clustering algorithms are discussed including hierarchical clustering, the K-means algorithm, fuzzy C-means algorithm, mixture density-based clustering, neural network-based clustering based on adaptive resonance theory, spectral clustering, subspace clustering and biclustering, and deep learning clustering.

In [8] unsupervised learning algorithms and latent variable models are presented. Basic linear and multilinear models for matrix and tensor factorizations and decompositions are discussed. Constrained matrix and tensor decompositions for sparse representation of data and their extensions are addressed. Various constraints such as orthogonality, statistical independence, nonnegativity and/or sparsity are explained. The importance of matrix/tensor decompositions is given for blind source separation, dimensionality reduction, pattern recognition, object detection, classification, multiway clustering, sparse representation and coding and data fusion.

In [9] an introduction is presented on semi-supervised learning. Discussed topics include transductive support vector machine and low density separation, co-training and multi-view, co-regularization and expectation-maximization for mixture models. Finally graph-based semi-supervised learning is addressed with graph Laplacian regularization, manifold regularization, measure-based regularization, and semi-supervised learning for structured outputs.

In [10] an overview is given on sparsity-aware learning and compressed sensing. The Least Absolute Shrinkage and Selection Operator (LASSO), sparse signal representation, ℓ_2, ℓ_0, ℓ_1 norm minimizers and their geometric interpretation are discussed. In view of conditions for equivalence of the ℓ_0 and ℓ_1 minimizer, mutual coherence and the Restricted Isometry Property (RIP) is explained. Robust sparse signal recovery from noisy measure-

ments and compressed sensing is covered. Sparsity-promoting algorithms are discussed like Orthogonal Matching Pursuit, the Least Angle Regression (LARS) algorithm and Iterative Shrinkage Algorithms. A case study on time-frequency analysis is provided.

In [11] the authors present information based learning approaches. Starting from information theoretic descriptors as entropy, divergence and mutual information, a unifying information theoretic framework for machine learning is outlined. Filtering, classification, feature extraction and nonparametric information estimators are discussed. Next a reproducing kernel Hilbert space framework for information based learning is proposed. Illustrative examples are given on adaptive system training, classification, information cut for clustering and independent component analysis.

In [12] model selection aspects are discussed. The Akaike information criterion and the Kullback information criterion are explained with linear regression as an example application. Then consistency and efficiency are addressed. Other topics that are included are Bayesian approaches to model selection, the Bayesian information criterion, Markov-Chain Monte-Carlo Bayesian methods, model selection by compression, minimum message length, model selection consistency, parameter estimation consistency and sequential variants of minimum description length.

In [13] an overview is given on music mining. Topics that are addressed include ground truth acquisition and evaluation, audio feature extraction, extracting context information about music, content-based similarity retrieval, genre classification, emotion/mood classification, music clustering, automatic tag annotation, audio fingerprinting and cover song detection.

Acknowledgments. The author acknowledges support from KU Leuven, the Flemish government, FWO, the Belgian federal science policy office and the European Research Council (ERC AdG A-DATADRIVE-B, CoE EF/05/006, GOA MANET, IUAP DYSCO, FWO G.0377.12, POM II, Cost IntelliCIS, iMinds Future health department).

References

[1] A. Tewari, P.L. Bartlett, Learning Theory, *Academic Press' Library in Signal Processing, Vol.1*, (Eds. S. Theodoridis, R. Chellappa), 2013.

[2] B. Hammer, Neural Networks, *Academic Press' Library in Signal Processing, Vol.1*, (Eds. S. Theodoridis, R. Chellappa), 2013.

[3] J. Shawe-Taylor, S. Sun, Kernel Methods and Support Vector Machines, *Academic Press' Library in Signal Processing, Vol.1*, (Eds. S. Theodoridis, R. Chellappa), 2013.

[4] K. Slavakis, P. Bouboulis, S. Theodoridis, Online Learning in Reproducing Kernel Hilbert Spaces, *Academic Press' Library in Signal Processing, Vol.1*, (Eds. S. Theodoridis, R. Chellappa), 2013.

[5] F. Pernkopf, R. Peharz, S. Tschiatschek, Introduction to Probabilistic Graphical Models, *Academic Press' Library in Signal Processing, Vol.1*, (Eds. S. Theodoridis, R. Chellappa), 2013.

[6] A.T. Cemgil, A Tutorial Introduction to Monte Carlo Methods, Markov Chain Monte Carlo and Particle Filtering, *Academic Press' Library in Signal Processing, Vol.1*, (Eds. S. Theodoridis, R. Chellappa), 2013.

[7] D. Lam, D.C. Wunsch, Clustering, *Academic Press' Library in Signal Processing, Vol.1*, (Eds. S. Theodoridis, R. Chellappa), 2013.

[8] A. Cichocki, Unsupervised Learning Algorithms and Latent Variable Models: PCA/SVD, CCA/PLS, ICA, NMF, etc., *Academic Press' Library in Signal Processing, Vol.1*, (Eds. S. Theodoridis, R. Chellappa), 2013.

[9] X. Zhou, M. Belkin, Semi-Supervised Learning, *Academic Press' Library in Signal Processing, Vol.1*, (Eds. S. Theodoridis, R. Chellappa), 2013.

[10] S. Theodoridis, Y. Kopsinis, K. Slavakis, Sparsity-Aware Learning and Compressed Sensing: An Overview, *Academic Press' Library in Signal Processing, Vol.1*, (Eds. S. Theodoridis, R. Chellappa), 2013.

[11] J.C. Principe, B. Chen, L.G. Sanchez Giraldo, Information Based Learning, *Academic Press' Library in Signal Processing, Vol.1*, (Eds. S. Theodoridis, R. Chellappa), 2013.

[12] E. Makalic, D.F. Schmidt, A.-K. Seghouane, A Tutorial on Model Selection, *Academic Press' Library in Signal Processing, Vol.1*, (Eds. S. Theodoridis, R. Chellappa), 2013.

[13] G. Tzanetakis, Music Mining, *Academic Press' Library in Signal Processing, Vol.1*, (Eds. S. Theodoridis, R. Chellappa), 2013.

[14] N. Aronszajn, "Theory of reproducing kernels", *Trans. American Mathematical Soc.*, **68**, 337-404, 1950.

[15] G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, S.V.N. Vishwanathan (Eds.), *Predicting Structured Data*, MIT Press, Cambridge, MA, 2007.

[16] R.G. Baraniuk, V. Cevher, M.F. Duarte, C. Hegde, "Model-based compressive sensing", *IEEE Transactions on Information Theory*, Vol. 56, No. 4, pp. 1982-2001, 2010.

[17] H.H. Bauschke, P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2011.

[18] M. Belkin, P. Niyogi, V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *Journal of Machine Learning Research*, 7: 2399-2434, 2006.

[19] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[20] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006

[21] S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[22] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers", *Foundations and Trends in Machine Learning*, 3(1): 1-122, 2011.

[23] A.M. Bruckstein, D.L. Donoho, M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images", *SIAM Review*, Vol. 51, No. 1, pp. 34-81, 2009.

[24] E.J. Candes, J. Romberg, T. Tao, "Stable recovery from incomplete and inaccurate measurements", *Communications on Pure and Applied Mathematics*, Vol. 59, No. 8, pp. 1207-1223, 2006.

[25] E.J. Candes, M.B. Wakin, "An introduction to compressive sampling", *IEEE Signal Processing Magazine*, 25(2): 21-30, 2008.

[26] O. Chapelle, B. Schölkopf, A. Zien (Eds.), *Semi-Supervised Learning*, MIT Press, 2006.

[27] S. Chen, D.L. Donoho, M. Saunders, "Atomic decomposition by basis pursuit", *SIAM Journal on Scientific Computing*, Vol. 20, No. 1, pp. 33-61, 1998.

[28] F.R.K. Chung, *Spectral graph theory*, in: CBMS Regional Conference Series in Mathematics, No. 92, 1992.

[29] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[30] F. Cucker, S. Smale, "On the mathematical foundations of learning theory", *Bulletin of the AMS*, 39, 1-49, 2002.

[31] F. Cucker, D.-X. Zhou, *Learning Theory: an Approximation Theory Viewpoint*, Cambridge University Press, 2007.

[32] D.L. Donoho, M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization", *Proceedings of National Academy of Sciences*, pp. 2197-2202, 2003.

[33] R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification* (2nd ed.), John Wiley & Sons, New York, 2001.

[34] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning,* Springer-Verlag, 2001.

[35] S. Haykin, *Adaptive Filter Theory*, Third Edition, Prentice-Hall, 1996.

[36] K. Hornik, M. Stinchcombe, H. White, "Multilayer feedforward networks are universal approximators", *Neural Networks*, Vol.2, pp.359-366, 1989.

[37] M.I. Jordan, *Learning in Graphical Models*, MIT Press, 1999.

[38] I.T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics, Springer-Verlag, 1986.

[39] T. Kailath, "RKHS approach to detection and estimation problems: Part I: deterministic signals in Gaussian noise", *IEEE Transactions on Information Theory*, **17**(5), 530-549, 1971.

[40] T. Kailath, "A view of three decades of linear filtering theory", *IEEE Transactions on Information Theory*, **20**(2), 146-181, 1974.

[41] D. Koller, N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 2009.

[42] W. Liu, J. Principe, S. Haykin, *Kernel Adaptive Filtering: A Comprehensive Introduction*, Wiley, Hoboken, New Jersey, 2010.

[43] D.J.C. MacKay, "Introduction to Gaussian processes". in *Neural networks and machine learning* (Ed. C.M. Bishop), Springer NATO-ASI Series F: Computer and Systems Sciences, Vol.168, 133-165, 1998.

[44] D.J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

[45] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations", *Philos. Trans. Roy. Soc. London*, **209**, 415-446, 1909.

[46] E.H. Moore, "On properly positive Hermitian matrices", *Bull. Amer. Math. Soc.*, **23**, 59, 1916.

[47] Y.E. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$", *Dokl. Akad. Nauk SSSR*, Vol. 269, pp. 543-547, 1983 (in Russian).

[48] E. Parzen, "Statistical inference on time series by RKHS methods", Dep. Statist. Stanford Univ. Tech. Rep.14, Jan. 1970.

[49] T. Poggio, F. Girosi, "Networks for approximation and learning", *Proceedings of the IEEE*, **78**(9), 1481-1497, 1990.

[50] J.C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, Springer, 2010.

[51] C.E. Rasmussen, C.K.I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[52] B. Schölkopf, A. Smola, K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem", *Neural Computation,* 10, 1299-1319, 1998.

[53] B. Schölkopf, A. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

[54] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, June 2004.

[55] I. Steinwart, A. Christmann, *Support Vector Machines*, New York: Springer, 2008.

[56] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, *Least Squares Support Vector Machines*, World Scientific, Singapore, 2002.

[57] J.A.K. Suykens, C. Alzate, K. Pelckmans, "Primal and dual model representations in kernel-based learning", *Statistics Surveys,* **4**, 148-183, 2010.

[58] S. Theodoridis, K. Koutroumbas, *Pattern Recognition*, Academic Press, 2009.

[59] A.N. Tikhonov, V.Y. Arsenin, *Solution of Ill-Posed Problems*, Winston, Washington DC, 1977.

[60] V. Vapnik, A. Lerner, "Pattern recognition using generalized portrait method", *Automation and Remote Control*, **24**, 774-780, 1963.

[61] V. Vapnik, A. Chervonenkis, "A note on one class of perceptrons", *Automation and Remote Control*, **25**, 1964.

[62] V. Vapnik, *Statistical Learning Theory,* John Wiley & Sons, New York, 1998.

[63] G. Wahba, *Spline Models for Observational Data*, Series in Applied Mathematics, 59, SIAM, Philadelphia, 1990.

[64] B. Widrow, R.G. Winter, "Neural Nets for Adaptive Filtering and Adaptive Pattern Recognition", *IEEE Computer Magazine*, 21(3):25-39, 1988.