Least Squares Support Vector Machines

This page is intentionally left blank

Least Squares Support Vector Machines

**Johan A. K. Suykens, Tony Van Gestel,
Jos De Brabanter, Bart De Moor and Joos Vandewalle**

K. U. Leuven, Belgium

World Scientific

New Jersey • London • Singapore • Hong Kong

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data

Least squares support vector machines / Johan A.K. Suykens ... [et al.].

p. cm.

Includes bibliographical references and index.

ISBN 9812381511 (alk. paper)

1. Machine learning. 2. Algorithms. 3. Kernel functions. 4. Least squares. I. Suykens, Johan A. K.

Q325.5 .L45 2002

006.3'1--dc21

2002033063

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

First published 2002

Reprinted 2005

Copyright © 2002 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

Printed by FuIsland Offset Printing (S) Pte Ltd, Singapore

Preface

In recent years there have been many new and exciting developments in kernel based learning, largely stimulated by work in statistical learning theory and support vector machines. Our research in this area started in 1998 around the time that we organized the *International Workshop on Advanced Black-Box Techniques for Nonlinear Modelling* in Leuven, where Vladimir Vapnik presented his important breakthroughs in this area. Solving nonlinear modelling and classification problems by convex optimization without suffering from many local minima sounded indeed very appealing and interesting for a deeper investigation.

Driven by the dream to make the approach as simple as possible (but not simpler) led us to the formulation of least squares support vector machine classifiers, as a first contribution in this area. Many tests and comparisons showed great performance of LS-SVMs on several benchmark data set problems and were very encouraging for further research in this promising direction. At the ESAT-SISTA research division of the Electrical Engineering department of the Katholieke Universiteit Leuven a lot of expertise in the area of mathematical engineering, including neural networks, was available which largely motivated the study of least squares support vector machines from this perspective. Conceptually, the additional explicit *primal-dual* interpretations from the viewpoint of optimization theory turned out to be essential for further developments of least squares support vector machines. In the neural networks area the emphasis has always been on *universal models* with applications within a very broad context such as function estimation, recurrent modelling, classification, control, unsupervised learning, on-line learning and many more. Links between LS-SVMs and regulariza-

LS-SVM: an interdisciplinary topic.

tion networks, Gaussian processes and kernel Fisher discriminant analysis became clear. A Bayesian learning and a robust statistics framework were developed. Support vector machine formulations to principal component analysis and canonical correlation analysis and their kernel versions were made. Towards large scale problems the primal-dual formulations have been exploited in a fixed size LS-SVM algorithm. It turned out that several extensions of the SVM methodology are much easier to formulate in terms of least squares and equality constraints instead of other loss functions or inequality constraints. Therefore, one of the important motivations for this book is to present a *general framework* (in the sense of traditional neural nets) for a class of support vector machines towards supervised and unsupervised learning and feedforward as well as recurrent networks.

The topic of least squares support vector machines is also very interdisciplinary. The world in which we are living is characterized by fragmentation, but at the same time the emergence of several new technologies requires different fields to interact with each other. Given these two sometimes conflicting faces of reality, it is important to have common languages for

transferring ideas between different fields and translating them into novel applications. Therefore another motivation for this book is to offer an *interdisciplinary forum* where different fields can meet, ranging from neural networks, machine learning, mathematics, statistics, optimization, pattern recognition, signal processing, circuits-, systems- and control theory to many applications areas. At this point we were stimulated by invitations and organizations of special sessions at international conferences and the opportunities for invited talks, tutorials and mini-courses such as for *IJCNN*, *ESANN*, *ECCTD*, *ISCAS*, *AMS*, *IMTC*, *ECC*, *FoCM*, *ICRM*. In parallel with the writing of this book, a *NATO Advanced Study Institute on Learning Theory and Practice* was organized in Leuven July 2002, which has been particularly motivating as well.

More specifically, for stimulating discussions, invitations and organization of joint meetings, we would like to thank Peter Bartlett, Sankar Basu, Jan Beirlant, Colin Campbell, Vladimir Cherkassky, Nello Cristianini, Felipe Cucker, Mark Embrechts, Lee Feldkamp, Martin Hasler, Simon Haykin, Gabor Horvath, Sathya Keerthi, Anthony Kuh, Chih-Jen Lin, Lennart Ljung, Charlie Micchelli, Tommy Poggio, Massi Pontil, Danil Prokhorov, Johan Schoukens, Steve Smale, Stefan Vandewalle, Paul Van Dooren, Vladimir Vapnik, Mathukumalli Vidyasagar, Grace Wahba, Paul Werbos, Yu Yi, and many others.

All of us also highly appreciated the great efforts made by Bernhard Schölkopf and Alex Smola for setting up a website on kernel machines www.kernel-machines.org in recent years. For least squares support vector machines a Matlab/C toolbox called LS-SVMLab is available at

<http://www.esat.kuleuven.ac.be/sista/lssvmlab/>.

We are especially grateful to Kristiaan Pelckmans for this development and to his colleagues Lukas, Bart Hamers and former master student Emmanuel Lambert.

Besides theoretical and algorithmical contributions on least squares support vector machines, several applications studies have been made with joint projects at K.U. Leuven. In the area of *bio-informatics* with applications to microarray data and textmining, we are grateful to Peter Antal, Tijl Debie, Frank De Smet, Patrick Glenisson, Bart Hamers, Kathleen Marchal, Janick Mathys, Yves Moreau and Gert Thijs. For *biomedical applications* we enjoyed our cooperation with Sabine Van Huffel in the projects

on classification of brain tumours from magnetic resonance spectroscopy signals (with Andy Devos, Lukas, Rene Intzand, Leentje Vanhamme, and Rosemary Tate (University of Sussex)), detection of ovarian cancer (with Chuan Lu, and Dirk Timmerman, Ignace Vergote (K.U. Leuven Hospitals)) and prediction of mental development of preterm newborns (with Lieveke Ameeye, and Hans Daniels, Gunnar Naulaers, Hugo Devlieger (K.U. Leuven Hospitals)). In the area of *nonlinear system identification and control* we are grateful to Luc Hoegaerts, Jeroen Buijs, several master students and Jakobus Barnard, Chris Aldrich (University of Stellenbosh) for our cooperation on the problem of prediction of air pollution. For *benchmarking studies and marketing applications* we want to thank Bart Baesens, Stijn Viaene, Jan Vanthienen, Guido Dedene (K.U. Leuven Applied Economic Sciences). For studies in *financial engineering* we are grateful to Dirk Baestaens (Fortis Bank Brussels) and several master students, in particular also Gert Lanckriet for contributions to the Bayesian LS-SVM framework.

Furthermore we are grateful to our many colleagues at other universities within the interuniversity poles of attraction in Belgium, the interdisciplinary center of neural networks of the K.U. Leuven, all the members of our (continuously growing) research group ESAT-SCD-SISTA and the interaction with its spin-off companies. Thanks to all of you for the great atmosphere!

Neural networks have often been presented as a universal solution to many real-life problems (sometimes even as a “miracle solution”) and one has come in a stage now where it is important to understand the limits of intelligence, both artificial and human as stated by Steve Smale in his *mathematical problems for the next century* [218] (Problem 18). We are very grateful to him and others as Tommy Poggio and Felipe Cucker for invitations to special meetings on learning theory and hope that this book, together with several breakthroughs in this area, may further contribute towards understanding these exciting problems.

Johan Suykens
 Tony Van Gestel
 Jos De Brabanter
 Bart De Moor
 Joos Vandewalle
Leuven, June 2002

Acknowledgements

Our research is supported by grants from several funding agencies and sources: Research Council K.U. Leuven: Concerted Research Action GOA-Mefisto 666 (Mathematical Engineering), IDO (IOTA Oncology, Genetic networks), several PhD/postdoc & fellow grants; Flemish Government: Fund for Scientific Research FWO Flanders (several PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0080.01 (collective intelligence), G.0256.97 (subspace), G.0115.01 (bio-i and microarrays), G.0240.99 (multilinear algebra), G.0197.02 (power islands), research communities IC-CoS, ANMMM), AWI (Bil. Int. Collaboration South Africa, Hungary and Poland), IWT (Soft4s (softsensors), STWW-Genprom (gene promotor prediction), GBOU McKnow (Knowledge management algorithms), Eureka-Impact (MPC-control), Eureka-FLiTE (flutter modeling), several PhD-grants); Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006): Dynamical Systems and Control: Computation, Identification & Modelling), Program Sustainable Development PODO-II (CP-TR-18: Sustainability effects of Traffic Management Systems); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS. JS is a professor at K.U. Leuven Belgium and a postdoctoral researcher with the FWO Flanders. TVG has been a research assistant with the FWO Flanders and is presently a postdoctoral researcher with the FWO Flanders. BDM and JVDW are full professors at K.U. Leuven Belgium.

This page is intentionally left blank

Contents

Chapter 1	Introduction	1
1.1	Multilayer perceptron neural networks	1
1.2	Regression and classification	4
1.3	Learning and generalization	7
1.3.1	Weight decay and effective number of parameters	7
1.3.2	Ridge regression	9
1.3.3	Bayesian learning	11
1.4	Principles of pattern recognition	14
1.4.1	Bayes rule and optimal classifier under Gaussian assumptions	14
1.4.2	Receiver operating characteristic	19
1.5	Dimensionality reduction methods	20
1.6	Parametric versus non-parametric approaches and RBF networks	22
1.7	Feedforward versus recurrent network models	24
Chapter 2	Support Vector Machines	29
2.1	Maximal margin classification and linear SVMs	29
2.1.1	Margin	29
2.1.2	Linear SVM classifier: separable case	30
2.1.3	Linear SVM classifier: non-separable case	34
2.2	Kernel trick and Mercer condition	36
2.3	Nonlinear SVM classifiers	37
2.4	VC theory and structural risk minimization	44

2.4.1 Empirical risk versus generalization error 44

2.4.2 Structural risk minimization 48

2.5 SVMs for function estimation 51

2.5.1 SVM for linear function estimation 51

2.5.2 SVM for nonlinear function estimation 54

2.5.3 VC bound on generalization error 56

2.6 Modifications and extensions 60

2.6.1 Kernels 60

2.6.2 Extension to other convex cost functions 63

2.6.3 Algorithms 65

2.6.4 Parametric versus non-parametric approaches 68

**Chapter 3 Basic Methods of Least Squares Support Vector
Machines 71**

3.1 Least Squares Support Vector Machines for classification 71

3.2 Multi-class formulations 76

3.3 Link with Fisher discriminant analysis in feature space 81

3.3.1 Linear Fisher discriminant analysis 82

3.3.2 Fisher discriminant analysis in feature space 85

3.4 Solving the LS-SVM KKT system 86

3.4.1 Conjugate gradient iterative methods 86

3.4.2 LS-SVM classifiers: UCI benchmarking results 89

3.5 Least Squares Support Vector Machines for function estimation 98

3.6 Links with regularization networks and Gaussian processes . . 100

3.6.1 RKHS and reproducing property 101

3.6.2 Representer theorem and regularization networks 104

3.6.3 Gaussian processes 106

3.7 Sparseness by pruning 111

Chapter 4 Bayesian Inference for LS-SVM Models 117

4.1 Bayesian inference for LS-SVM classifiers 117

4.1.1 Definition of network model, parameters and hyperpa-
rameters 117

4.1.2 Bayes rule and levels of inference 119

4.1.3 Probabilistic interpretation of LS-SVM classifiers
(Level 1) 122

4.1.4 Inference of the hyperparameters (Level 2) 129

4.1.5	Inference of kernel parameters and model comparison (Level 3)	131
4.1.6	Design of the LS-SVM classifier in the Bayesian evidence framework	134
4.1.7	Example and benchmarking results	136
4.2	Bayesian inference for LS-SVM regression	139
4.2.1	Probabilistic interpretation of LS-SVM regressors (Level 1): predictive mean and error bars	140
4.2.2	Inference of hyperparameters (Level 2)	143
4.2.3	Inference of kernel parameters and model comparison (Level 3)	144
4.3	Input selection by automatic relevance determination	144
Chapter 5 Robustness		149
5.1	Noise model assumptions and robust statistics	149
5.2	Weighted LS-SVMs	150
5.3	Robust cross-validation	156
5.3.1	M -Estimators	158
5.3.2	L -Estimators	163
5.3.3	Efficiency-robustness trade-off	165
5.3.4	A robust and efficient cross-validation score function	167
Chapter 6 Large Scale Problems		173
6.1	Low rank approximation methods	173
6.1.1	Nyström method	173
6.1.2	Incomplete Cholesky factorization	176
6.2	Fixed Size LS-SVMs	178
6.2.1	Estimation in primal weight space	179
6.2.2	Active selection of support vectors	181
6.3	Basis construction in the feature space	183
6.4	Combining submodels	190
6.4.1	Committee network approach	190
6.4.2	Multilayer networks of LS-SVMs	196
Chapter 7 LS-SVM for Unsupervised Learning		201
7.1	Support Vector Machines and linear PCA analysis	201
7.1.1	Classical principal component analysis formulation	202
7.1.2	Support vector machine formulation to linear PCA	202

7.1.3 Including a bias term 205

7.1.4 The reconstruction problem 208

7.2 An LS-SVM approach to kernel PCA 210

7.3 Links with density estimation 214

7.4 Kernel CCA 219

7.4.1 Classical canonical correlation analysis formulation . . . 219

7.4.2 Support vector machine formulation to linear CCA . . . 220

7.4.3 Extension to kernel CCA 222

Chapter 8 LS-SVM for Recurrent Networks and Control 225

8.1 Recurrent Least Squares Support Vector Machines 225

8.1.1 From feedforward to recurrent LS-SVM formulations . . 225

8.1.2 A simplification to the problem 228

8.1.3 Example: trajectory learning of chaotic systems 230

8.2 LS-SVMs and optimal control 232

8.2.1 The N -stage optimal control problem 232

8.2.2 LS-SVMs as controllers within the N -stage optimal control problem 234

8.2.3 Alternative formulation and stability issues 238

8.2.4 Illustrative examples 239

Appendix A 249

Bibliography 269

List of Symbols 287

Acronyms 289

Index 291