

Contents

Preface	v
Organizing committee	ix
List of chapter contributors	xi
1 An Overview of Statistical Learning Theory	1
<i>V. Vapnik</i>	
1.1 Setting of the Learning Problem	2
1.1.1 Function estimation model	2
1.1.2 Problem of risk minimization	2
1.1.3 Three main learning problems	2
1.1.4 Empirical risk minimization induction principle	4
1.1.5 Empirical risk minimization principle and the classical methods	4
1.1.6 Four parts of learning theory	5
1.2 The Theory of Consistency of Learning Processes	6
1.2.1 The key theorem of the learning theory	6
1.2.2 The necessary and sufficient conditions for uniform convergence	7
1.2.3 Three milestones in learning theory	9
1.3 Bounds on the Rate of Convergence of the Learning Processes	10
1.3.1 The structure of the growth function	11
1.3.2 Equivalent definition of the VC dimension	11
1.3.3 Two important examples	12
1.3.4 Distribution independent bounds for the rate of convergence of learning processes	13
1.3.5 Problem of constructing rigorous (distribution dependent) bounds	14
1.4 Theory for Controlling the Generalization of Learning Machines	15
1.4.1 Structural risk minimization induction principle	15
1.5 Theory of Constructing Learning Algorithms	17
1.5.1 Methods of separating hyperplanes and their generalization . . .	17
1.5.2 Sigmoid approximation of indicator functions and neural nets .	18
1.5.3 The optimal separating hyperplanes	19
1.5.4 The support vector network	21
1.5.5 Why can neural networks and support vectors networks generalize?	23
1.6 Conclusion	24

2	Best Choices for Regularization Parameters in Learning Theory: On the Bias-Variance Problem	29
	<i>F. Cucker, S. Smale</i>	
2.1	Introduction	30
2.2	RKHS and Regularization Parameters	30
2.3	Estimating the Confidence	32
2.4	Estimating the Sample Error	38
2.5	Choosing the optimal γ	40
2.6	Final Remarks	41
3	Cucker Smale Learning Theory in Besov Spaces	47
	<i>C.A. Micchelli, Y. Xu, P. Ye</i>	
3.1	Introduction	48
3.2	Cucker Smale Functional and the Peetre K-Functional	48
3.3	Estimates for the CS-Functional in Anisotropic Besov Spaces	52
4	High-dimensional Approximation by Neural Networks	69
	<i>V. Kůrková</i>	
4.1	Introduction	70
4.2	Variable-basis Approximation and Optimization	71
4.3	Maurey-Jones-Barron's Theorem	73
4.4	Variation with respect to a Set of Functions	75
4.5	Rates of Approximate Optimization over Variable Basis Functions	77
4.6	Comparison with Linear Approximation	79
4.7	Upper Bounds on Variation	80
4.8	Lower Bounds on Variation	82
4.9	Rates of Approximation of Real-valued Boolean Functions	83
5	Functional Learning through Kernels	89
	<i>S. Canu, X. Mary, A. Rakotomamonjy</i>	
5.1	Some Questions Regarding Machine Learning	90
5.2	<i>r.k.h.s</i> Perspective	91
5.2.1	Positive kernels	91
5.2.2	<i>r.k.h.s</i> and learning in the literature	91
5.3	Three Principles on the Nature of the Hypothesis Set	92
5.3.1	The learning problem	92
5.3.2	The evaluation functional	93
5.3.3	Continuity of the evaluation functional	93
5.3.4	Important consequence	94
5.3.5	$\mathbb{R}^{\mathcal{X}}$ the set of the pointwise defined functions on \mathcal{X}	94
5.4	Reproducing Kernel Hilbert Space (<i>r.k.h.s</i>)	95
5.5	Kernel and Kernel Operator	97
5.5.1	How to build <i>r.k.h.s</i> ?	97
5.5.2	Carleman operator and the regularization operator	98
5.5.3	Generalization	99
5.6	Reproducing Kernel Spaces (<i>r.k.k.s</i>)	99

5.6.1	Evaluation spaces	99
5.6.2	Reproducing kernels	100
5.7	Representer Theorem	104
5.8	Examples	105
5.8.1	Examples in Hilbert space	105
5.8.2	Other examples	107
5.9	Conclusion	107
6	Leave-one-out Error and Stability of Learning Algorithms with Ap- plications	111
	<i>A. Elisseeff, M. Pontil</i>	
6.1	Introduction	112
6.2	General Observations about the Leave-one-out Error	113
6.3	Theoretical Attempts to Justify the Use of the Leave-one-out Error	116
6.3.1	Early work in non-parametric statistics	116
6.3.2	Relation to VC-theory	117
6.3.3	Stability	118
6.3.4	Stability of averaging techniques	119
6.4	Kernel Machines	119
6.4.1	Background on kernel machines	120
6.4.2	Leave-one-out error for the square loss	121
6.4.3	Bounds on the leave-one-out error and stability	122
6.5	The Use of the Leave-one-out Error in Other Learning Problems	123
6.5.1	Transduction	123
6.5.2	Feature selection and rescaling	123
6.6	Discussion	124
6.6.1	Sensitivity analysis, stability, and learning	124
6.6.2	Open problems	124
7	Regularized Least-Squares Classification	131
	<i>R. Rifkin, G. Yeo, T. Poggio</i>	
7.1	Introduction	132
7.2	The RLSC Algorithm	134
7.3	Previous Work	135
7.4	RLSC vs. SVM	136
7.5	Empirical Performance of RLSC	137
7.6	Approximations to the RLSC Algorithm	139
7.6.1	Low-rank approximations for RLSC	141
7.6.2	Nonlinear RLSC application: image classification	142
7.7	Leave-one-out Bounds for RLSC	146
8	Support Vector Machines: Least Squares Approaches and Extensions	155
	<i>J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle</i>	
8.1	Introduction	156
8.2	Least Squares SVMs for Classification and Function Estimation	158
8.2.1	LS-SVM classifiers and link with kernel FDA	158

8.2.2	Function estimation case and equivalence to a regularization network solution	161
8.2.3	Issues of sparseness and robustness	161
8.2.4	Bayesian inference of LS-SVMs and Gaussian processes	163
8.3	Primal-dual Formulations to Kernel PCA and CCA	163
8.3.1	Kernel PCA as a one-class modelling problem and a primal-dual derivation	163
8.3.2	A support vector machine formulation to Kernel CCA	166
8.4	Large Scale Methods and On-line Learning	168
8.4.1	Nyström method	168
8.4.2	Basis construction in the feature space using fixed size LS-SVM	169
8.5	Recurrent Networks and Control	172
8.6	Conclusions	173
9	Extension of the ν-SVM Range for Classification	179
	<i>F. Pérez-Cruz, J. Weston, D.J.L. Herrmann, B. Schölkopf</i>	
9.1	Introduction	180
9.2	ν Support Vector Classifiers	181
9.3	Limitation in the Range of ν	185
9.4	Negative Margin Minimization	186
9.5	Extended ν -SVM	188
9.5.1	Kernelization in the dual	189
9.5.2	Kernelization in the primal	191
9.6	Experiments	191
9.7	Conclusions and Further Work	194
10	Kernels Methods for Text Processing	197
	<i>N. Cristianini, J. Kandola, A. Vinokourov, J. Shawe-Taylor</i>	
10.1	Introduction	198
10.2	Overview of Kernel Methods	198
10.3	From Bag of Words to Semantic Space	199
10.4	Vector Space Representations	201
10.4.1	Basic vector space model	203
10.4.2	Generalised vector space model	204
10.4.3	Semantic smoothing for vector space models	204
10.4.4	Latent semantic kernels	205
10.4.5	Semantic diffusion kernels	207
10.5	Learning Semantics from Cross Language Correlations	211
10.6	Hypertext	215
10.7	String Matching Kernels	216
10.7.1	Efficient computation of SSK	219
10.7.2	n -grams- a language independent approach	220
10.8	Conclusions	220

11 An Optimization Perspective on Kernel Partial Least Squares Regression	227
<i>K.P. Bennett, M.J. Embrechts</i>	
11.1 Introduction	228
11.2 PLS Derivation	229
11.2.1 PCA regression review	229
11.2.2 PLS analysis	231
11.2.3 Linear PLS	232
11.2.4 Final regression components	234
11.3 Nonlinear PLS via Kernels	236
11.3.1 Feature space K-PLS	236
11.3.2 Direct kernel partial least squares	237
11.4 Computational Issues in K-PLS	238
11.5 Comparison of Kernel Regression Methods	239
11.5.1 Methods	239
11.5.2 Benchmark cases	240
11.5.3 Data preparation and parameter tuning	240
11.5.4 Results and discussion	241
11.6 Case Study for Classification with Uneven Classes	243
11.7 Feature Selection with K-PLS	243
11.8 Thoughts and Conclusions	245
12 Multiclass Learning with Output Codes	251
<i>Y. Singer</i>	
12.1 Introduction	252
12.2 Margin-based Learning Algorithms	253
12.3 Output Coding for Multiclass Problems	257
12.4 Training Error Bounds	260
12.5 Finding Good Output Codes	262
12.6 Conclusions	263
13 Bayesian Regression and Classification	267
<i>C.M. Bishop, M.E. Tipping</i>	
13.1 Introduction	268
13.1.1 Least squares regression	268
13.1.2 Regularization	269
13.1.3 Probabilistic models	269
13.1.4 Bayesian regression	271
13.2 Support Vector Machines	272
13.3 The Relevance Vector Machine	273
13.3.1 Model specification	273
13.3.2 The effective prior	275
13.3.3 Inference	276
13.3.4 Making predictions	277
13.3.5 Properties of the marginal likelihood	278
13.3.6 Hyperparameter optimization	279

13.3.7	Relevance vector machines for classification	280
13.4	The Relevance Vector Machine in Action	281
13.4.1	Illustrative synthetic data: regression	281
13.4.2	Illustrative synthetic data: classification	283
13.4.3	Benchmark results	284
13.5	Discussion	285
14	Bayesian Field Theory: from Likelihood Fields to Hyperfields	289
	<i>J. Lemm</i>	
14.1	Introduction	290
14.2	The Bayesian framework	290
14.2.1	The basic probabilistic model	290
14.2.2	Bayesian decision theory and predictive density	291
14.2.3	Bayes' theorem: from prior and likelihood to the posterior	293
14.3	Likelihood models	295
14.3.1	Log-probabilities, energies, and density estimation	295
14.3.2	Regression	297
14.3.3	Inverse quantum theory	298
14.4	Prior models	299
14.4.1	Gaussian prior factors and approximate symmetries	299
14.4.2	Hyperparameters and hyperfields	303
14.4.3	Hyperpriors for hyperfields	308
14.4.4	Auxiliary fields	309
14.5	Summary	312
15	Bayesian Smoothing and Information Geometry	319
	<i>R. Kulhavý</i>	
15.1	Introduction	320
15.2	Problem Statement	321
15.3	Probability-Based Inference	322
15.4	Information-Based Inference	324
15.5	Single-Case Geometry	327
15.6	Average-Case Geometry	331
15.7	Similar-Case Modeling	332
15.8	Locally Weighted Geometry	336
15.9	Concluding Remarks	337
16	Nonparametric Prediction	341
	<i>L. Györfi, D. Schäfer</i>	
16.1	Introduction	342
16.2	Prediction for Squared Error	342
16.3	Prediction for 0 – 1 Loss: Pattern Recognition	346
16.4	Prediction for Log Utility: Portfolio Selection	348

17 Recent Advances in Statistical Learning Theory	357
<i>M. Vidyasagar</i>	
17.1 Introduction	358
17.2 Problem Formulations	358
17.2.1 Uniform convergence of empirical means	358
17.2.2 Probably approximately correct learning	360
17.3 Summary of “Classical” Results	362
17.3.1 Fixed distribution case	362
17.3.2 Distribution-free case	364
17.4 Recent Advances	365
17.4.1 Intermediate families of probability measures	365
17.4.2 Learning with prior information	366
17.5 Learning with Dependent Inputs	367
17.5.1 Problem formulations	367
17.5.2 Definition of β -mixing	368
17.5.3 UCEM and PAC learning with β -mixing inputs	369
17.6 Applications to Learning with Inputs Generated by a Markov Chain . .	371
17.7 Conclusions	372
18 Neural Networks in Measurement Systems (an engineering view)	375
<i>G. Horváth</i>	
18.1 Introduction	376
18.2 Measurement and Modeling	377
18.3 Neural Networks	383
18.4 Support Vector Machines	389
18.5 The Nature of Knowledge, Prior Information	393
18.6 Questions Concerning Implementation	394
18.7 Conclusions	396
List of participants	403
Index	411