

Contents

Preface to the Second Edition	vii
Preface to the First Edition	ix
Introduction: Four Periods in the Research of the Learning Problem	1
Rosenblatt's Perceptron (The 1960s)	1
Construction of the Fundamentals of Learning Theory (The 1960s–1970s)	7
Neural Networks (The 1980s)	11
Returning to the Origin (The 1990s)	14
Chapter 1 Setting of the Learning Problem	17
1.1 Function Estimation Model	17
1.2 The Problem of Risk Minimization	18
1.3 Three Main Learning Problems	18
1.3.1 Pattern Recognition	19
1.3.2 Regression Estimation	19
1.3.3 Density Estimation (Fisher–Wald Setting)	19
1.4 The General Setting of the Learning Problem	20
1.5 The Empirical Risk Minimization (ERM) Inductive Principle	20
1.6 The Four Parts of Learning Theory	21
Informal Reasoning and Comments — 1	23

1.7	The Classical Paradigm of Solving Learning Problems . . .	23
1.7.1	Density Estimation Problem (Maximum Likelihood Method)	24
1.7.2	Pattern Recognition (Discriminant Analysis) Problem	24
1.7.3	Regression Estimation Model	25
1.7.4	Narrowness of the ML Method	26
1.8	Nonparametric Methods of Density Estimation	27
1.8.1	Parzen's Windows	27
1.8.2	The Problem of Density Estimation Is Ill-Posed . . .	28
1.9	Main Principle for Solving Problems Using a Restricted Amount of Information	30
1.10	Model Minimization of the Risk Based on Empirical Data .	31
1.10.1	Pattern Recognition	31
1.10.2	Regression Estimation	31
1.10.3	Density Estimation	32
1.11	Stochastic Approximation Inference	33
Chapter 2 Consistency of Learning Processes		35
2.1	The Classical Definition of Consistency and the Concept of Nontrivial Consistency	36
2.2	The Key Theorem of Learning Theory	38
2.2.1	Remark on the ML Method	39
2.3	Necessary and Sufficient Conditions for Uniform Two-Sided Convergence	40
2.3.1	Remark on Law of Large Numbers and Its Generalization	41
2.3.2	Entropy of the Set of Indicator Functions	42
2.3.3	Entropy of the Set of Real Functions	43
2.3.4	Conditions for Uniform Two-Sided Convergence . . .	45
2.4	Necessary and Sufficient Conditions for Uniform One-Sided Convergence	45
2.5	Theory of Nonfalsifiability	47
2.5.1	Kant's Problem of Demarcation and Popper's Theory of Nonfalsifiability	47
2.6	Theorems on Nonfalsifiability	49
2.6.1	Case of Complete (Popper's) Nonfalsifiability	50
2.6.2	Theorem on Partial Nonfalsifiability	50
2.6.3	Theorem on Potential Nonfalsifiability	52
2.7	Three Milestones in Learning Theory	55
Informal Reasoning and Comments — 2		59
2.8	The Basic Problems of Probability Theory and Statistics . .	60
2.8.1	Axioms of Probability Theory	60
2.9	Two Modes of Estimating a Probability Measure	63

2.10 Strong Mode Estimation of Probability Measures and the Density Estimation Problem	65
2.11 The Glivenko–Cantelli Theorem and its Generalization . . .	66
2.12 Mathematical Theory of Induction	67

Chapter 3 Bounds on the Rate of Convergence of Learning Processes 69

3.1 The Basic Inequalities	70
3.2 Generalization for the Set of Real Functions	72
3.3 The Main Distribution–Independent Bounds	75
3.4 Bounds on the Generalization Ability of Learning Machines	76
3.5 The Structure of the Growth Function	78
3.6 The VC Dimension of a Set of Functions	80
3.7 Constructive Distribution–Independent Bounds	83
3.8 The Problem of Constructing Rigorous (Distribution–Dependent) Bounds	85

Informal Reasoning and Comments — 3 87

3.9 Kolmogorov–Smirnov Distributions	87
3.10 Racing for the Constant	89
3.11 Bounds on Empirical Processes	90

Chapter 4 Controlling the Generalization Ability of Learning Processes 93

4.1 Structural Risk Minimization (SRM) Inductive Principle . .	94
4.2 Asymptotic Analysis of the Rate of Convergence	97
4.3 The Problem of Function Approximation in Learning Theory	99
4.4 Examples of Structures for Neural Nets	101
4.5 The Problem of Local Function Estimation	103
4.6 The Minimum Description Length (MDL) and SRM Principles	104
4.6.1 The MDL Principle	106
4.6.2 Bounds for the MDL Principle	107
4.6.3 The SRM and MDL Principles	108
4.6.4 A Weak Point of the MDL Principle	110

Informal Reasoning and Comments — 4 111

4.7 Methods for Solving Ill-Posed Problems	112
4.8 Stochastic Ill-Posed Problems and the Problem of Density Estimation	113
4.9 The Problem of Polynomial Approximation of the Regression	115
4.10 The Problem of Capacity Control	116
4.10.1 Choosing the Degree of the Polynomial	116
4.10.2 Choosing the Best Sparse Algebraic Polynomial . . .	117
4.10.3 Structures on the Set of Trigonometric Polynomials	118

- 4.10.4 The Problem of Features Selection 119
- 4.11 The Problem of Capacity Control and Bayesian Inference . 119
 - 4.11.1 The Bayesian Approach in Learning Theory 119
 - 4.11.2 Discussion of the Bayesian Approach and Capacity Control Methods 121

Chapter 5 Methods of Pattern Recognition 123

- 5.1 Why Can Learning Machines Generalize? 123
- 5.2 Sigmoid Approximation of Indicator Functions 125
- 5.3 Neural Networks 126
 - 5.3.1 The Back-Propagation Method 126
 - 5.3.2 The Back-Propagation Algorithm 130
 - 5.3.3 Neural Networks for the Regression Estimation Problem 130
 - 5.3.4 Remarks on the Back-Propagation Method 130
- 5.4 The Optimal Separating Hyperplane 131
 - 5.4.1 The Optimal Hyperplane 131
 - 5.4.2 Δ -margin hyperplanes 132
- 5.5 Constructing the Optimal Hyperplane 133
 - 5.5.1 Generalization for the Nonseparable Case 136
- 5.6 Support Vector (SV) Machines 138
 - 5.6.1 Generalization in High-Dimensional Space 139
 - 5.6.2 Convolution of the Inner Product 140
 - 5.6.3 Constructing SV Machines 141
 - 5.6.4 Examples of SV Machines 141
- 5.7 Experiments with SV Machines 146
 - 5.7.1 Example in the Plane 146
 - 5.7.2 Handwritten Digit Recognition 147
 - 5.7.3 Some Important Details 151
- 5.8 Remarks on SV Machines 154
- 5.9 SVM and Logistic Regression 156
 - 5.9.1 Logistic Regression 156
 - 5.9.2 The Risk Function for SVM 159
 - 5.9.3 The SVM_n Approximation of the Logistic Regression 160
- 5.10. Ensemble of the SVM 163
 - 5.10.1 The AdaBoost Method 164
 - 5.10.2 The Ensemble of SVMs 167

Informal Reasoning and Comments — 5 171

- 5.11 The Art of Engineering Versus Formal Inference 171
- 5.12 Wisdom of Statistical Models 174
- 5.13 What Can One Learn from Digit Recognition Experiments? 176
 - 5.13.1 Influence of the Type of Structures and Accuracy of Capacity Control 177

5.13.2	SRM Principle and the Problem of Feature Construction	178
5.13.3	Is the Set of Support Vectors a Robust Characteristic of the Data?	179
Chapter 6 Methods of Function Estimation		181
6.1	ε -Insensitive Loss-Function	181
6.2	SVM for Estimating Regression Function	183
6.2.1	SV Machine with Convolved Inner Product	186
6.2.2	Solution for Nonlinear Loss Functions	188
6.2.3	Linear Optimization Method	190
6.3	Constructing Kernels for Estimating Real-Valued Functions	190
6.3.1	Kernels Generating Expansion on Orthogonal Polynomials	191
6.3.2	Constructing Multidimensional Kernels	193
6.4	Kernels Generating Splines	194
6.4.1	Spline of Order d With a Finite Number of Nodes . .	194
6.4.2	Kernels Generating Splines With an Infinite Number of Nodes	195
6.5	Kernels Generating Fourier Expansions	196
6.5.1	Kernels for Regularized Fourier Expansions	197
6.6	The Support Vector ANOVA Decomposition for Function Approximation and Regression Estimation	198
6.7	SVM for Solving Linear Operator Equations	200
6.7.1	The Support Vector Method	201
6.8	Function Approximation Using the SVM	204
6.8.1	Why Does the Value of ε Control the Number of Support Vectors?	205
6.9	SVM for Regression Estimation	208
6.9.1	Problem of Data Smoothing	209
6.9.2	Estimation of Linear Regression Functions	209
6.9.3	Estimation Nonlinear Regression Functions	216
Informal Reasoning and Comments — 6		219
6.10	Loss Functions for the Regression Estimation Problem	219
6.11	Loss Functions for Robust Estimators	221
6.12	Support Vector Regression Machine	223
Chapter 7 Direct Methods in Statistical Learning Theory		225
7.1	Problem of Estimating Densities, Conditional Probabilities, and Conditional Densities	226
7.1.1	Problem of Density Estimation: Direct Setting	226
7.1.2	Problem of Conditional Probability Estimation	227
7.1.3	Problem of Conditional Density Estimation	228

7.2	Solving an Approximately Determined Integral Equation . . .	229
7.3	Glivenko-Cantelli Theorem	230
7.3.1	Kolmogorov-Smirnov Distribution	232
7.4	Ill-Posed Problems	233
7.5	Three Methods of Solving Ill-Posed Problems	235
7.5.1	The Residual Principle	236
7.6	Main Assertions of the Theory of Ill-Posed Problems	237
7.6.1	Deterministic Ill-Posed Problems	237
7.6.2	Stochastic Ill-Posed Problem	238
7.7	Nonparametric Methods of Density Estimation	240
7.7.1	Consistency of the Solution of the Density Estimation Problem	240
7.7.2	The Parzen's Estimators	241
7.8	SVM Solution of the Density Estimation Problem	244
7.8.1	The SVM Density Estimate: Summary	247
7.8.2	Comparison of the Parzen's and the SVM methods	248
7.9	Conditional Probability Estimation	249
7.9.1	Approximately Defined Operator	251
7.9.2	SVM Method for Conditional Probability Estimation	253
7.9.3	The SVM Conditional Probability Estimate: Summary	255
7.10	Estimation of Conditional Density and Regression	256
7.11	Remarks	258
7.11.1	One Can Use a Good Estimate of the Unknown Density	258
7.11.2	One Can Use Both Labeled (Training) and Unlabeled (Test) Data	259
7.11.3	Method for Obtaining Sparse Solutions of the Ill- Posed Problems	259
Informal Reasoning and Comments — 7		261
7.12	Three Elements of a Scientific Theory	261
7.12.1	Problem of Density Estimation	262
7.12.2	Theory of Ill-Posed Problems	262
7.13	Stochastic Ill-Posed Problems	263
Chapter 8 The Vicinal Risk Minimization Principle and the SVMs		267
8.1	The Vicinal Risk Minimization Principle	267
8.1.1	Hard Vicinity Function	269
8.1.2	Soft Vicinity Function	270
8.2	VRM Method for the Pattern Recognition Problem	271
8.3	Examples of Vicinal Kernels	275
8.3.1	Hard Vicinity Functions	276
8.3.2	Soft Vicinity Functions	279

8.4 Nonsymmetric Vicinities	279
8.5. Generalization for Estimation Real-Valued Functions	281
8.6 Estimating Density and Conditional Density	284
8.6.1 Estimating a Density Function	284
8.6.2 Estimating a Conditional Probability Function	285
8.6.3 Estimating a Conditional Density Function	286
8.6.4 Estimating a Regression Function	287
Informal Reasoning and Comments — 8	289
Chapter 9 Conclusion: What Is Important in Learning Theory?	291
9.1 What Is Important in the Setting of the Problem?	291
9.2 What Is Important in the Theory of Consistency of Learning Processes?	294
9.3 What Is Important in the Theory of Bounds?	295
9.4 What Is Important in the Theory for Controlling the Generalization Ability of Learning Machines?	296
9.5 What Is Important in the Theory for Constructing Learning Algorithms?	297
9.6 What Is the Most Important?	298
References	301
Remarks on References	301
References	302
Index	311