Communications and Control Engineering

M. Vidyasagar

Learning and Generalization

With Applications to Neural Networks

Second Edition

Springer

Table of Contents

Preface to the Second Edition xiii

Preface to the First Edition xvii

1. Introduction 1

2. Preliminaries 13
2.1 Pseudometric Spaces, Packing and Covering Numbers 13
2.1.1 Pseudometric Spaces 13
2.1.2 Packing and Covering Numbers 14
2.1.3 Compact and Totally Bounded Sets 16
2.2 Probability Measures 17
2.2.1 Definition of a Probability Space 17
2.2.2 A Pseudometric Induced by a Probability Measure 18
2.2.3 A Metric on the Set of Probability Measures 19
2.2.4 Random Variables 21
2.2.5 Conditional Expectations 23
2.3 Large Deviation Type Inequalities 24
2.3.1 Chernoff Bounds 24
2.3.2 Chernoff-Okamoto Bound 26
2.3.3 Hoeffding's Inequality 26
2.4 Stochastic Processes, Almost Sure Convergence 29
2.4.1 Probability Measures on Infinite Cartesian Products 29
2.4.2 Stochastic Processes 29
2.4.3 The Borel-Cantelli Lemma and Almost Sure Convergence 30
2.5 Mixing Properties of Stochastic Processes 33
2.5.1 Definitions of Various Kinds of Mixing Coefficients 34
2.5.2 Inequalities for Mixing Processes 36

3. Problem Formulations 43
3.1 Uniform Convergence of Empirical Means 43
3.1.1 The UCEM Property 43
3.1.2 The UCEMUP Property 52

3.1.3 Extension to Dependent Input Sequences 54
3.2 Learning Concepts and Functions 55
3.2.1 Concept Learning 55
3.2.2 Function Learning 64
3.2.3 Extension to Dependent Input Sequences 65
3.2.4 Assumptions Underlying the Model of Learning 66
3.2.5 Alternate Notions of Learnability 70
3.3 Model-Free Learning 76
3.3.1 Problem Formulation 76
3.3.2 Relationship to the Uniform Convergence of Empirical Means 81
3.4 Preservation of UCEMUP and PAC Properties 83
3.4.1 Preservation of UCEMUP Property with Beta-Mixing Inputs 84
3.4.2 Law of Large Numbers Under Alpha-Mixing Inputs ... 89
3.4.3 Preservation of PAC Learning Property with Beta-Mixing Inputs 94
3.4.4 Preservation of PAC Learning Property with Beta-Mixing Inputs: Continued 95
3.4.5 Replacing $\mathcal{P}$ by its Closure 97
3.5 Markov Chains and Beta-Mixing 100
3.5.1 Geometric Ergodicity and Beta-Mixing 100
3.5.2 Beta-Mixing Properties of Markov Sequences 105
3.5.3 Mixing Properties of Hidden Markov Models 110

4. Vapnik-Chervonenkis, Pseudo- and Fat-Shattering Dimensions 115
4.1 Definitions 115
4.1.1 The Vapnik-Chervonenkis Dimension 115
4.1.2 The Pseudo-Dimension 120
4.1.3 The Fat-Shattering Dimension 122
4.2 Bounds on Growth Functions 123
4.2.1 Growth Functions of Collections of Sets 123
4.2.2 Bounds on Covering Numbers Based on the Pseudo-Dimension 128
4.2.3 Metric Entropy Bounds for Families of Functions 132
4.2.4 Bounds on Covering Numbers Based on the Fat-Shattering Dimension 139
4.3 Growth Functions of Iterated Families 141

5. Uniform Convergence of Empirical Means 149
5.1 Restatement of the Problems Under Study 149
5.2 Equivalence of the UCEM and ASCEM Properties 153
5.3 Main Theorems 155
5.4 Preliminary Lemmas 161

5.5 Theorem 5.1: Proof of Necessity 173
5.6 Theorem 5.1: Proof of Sufficiency 178
5.7 Proofs of the Remaining Theorems 190
5.8 Uniform Convergence Properties of Iterated Families 194
5.8.1 Boolean Operations on Collections of Sets 195
5.8.2 Uniformly Continuous Mappings on Families of Functions 196
5.8.3 Families of Loss Functions 200

6. Learning Under a Fixed Probability Measure 207
6.1 Introduction 207
6.2 UCEM Property Implies ASEC Learnability 209
6.3 Finite Metric Entropy Implies Learnability 216
6.4 Consistent Learnability 224
6.4.1 Consistent PAC Learnability 224
6.4.2 Consistent PUAC Learnability 226
6.5 Examples 230
6.6 Learnable Concept Classes Have Finite Metric Entropy 236
6.7 Model-Free Learning 242
6.7.1 A Sufficient Condition for Learnability 244
6.7.2 A Necessary Condition 248
6.8 Dependent Inputs 250
6.8.1 Finite Metric Entropy and Alpha-Mixing Input Sequences 250
6.8.2 Consistent Learnability and Beta-Mixing Input Sequences 251

7. Distribution-Free Learning 255
7.1 Uniform Convergence of Empirical Means 255
7.1.1 Function Classes 256
7.1.2 Concept Classes 258
7.1.3 Loss Functions 261
7.2 Function Learning 263
7.2.1 Finite P-Dimension Implies PAC and PUAC Learnability 264
7.2.2 Finite P-Dimension is not Necessary for PAC Learnability 267
7.3 Concept Learning 269
7.3.1 Improved Upper Bound for the Sample Complexity ... 269
7.3.2 A Universal Lower Bound for the Sample Complexity . 273
7.3.3 Learnability Implies Finite VC-Dimension 278
7.4 Learnability of Functions with a Finite Range 280

8. Learning Under an Intermediate Family of Probabilities .. 285
8.1 General Families of Probabilities 287
8.1.1 Uniform Convergence of Empirical Means 287
8.1.2 Function Learning 288
8.1.3 Concept Learning 292
8.2 Totally Bounded Families of Probabilities 297
8.3 Families of Probabilities with a Nonempty Interior 308

9. Alternate Models of Learning 311
9.1 Efficient Learning 312
9.1.1 Definition of Efficient Learnability 313
9.1.2 The Complexity of Finding a Consistent Hypothesis .. 317
9.2 Active Learning 326
9.2.1 Fixed-Distribution Learning 329
9.2.2 Distribution-Free Learning 332
9.3 Learning with Prior Information: Necessary and Sufficient Conditions 335
9.3.1 Definition of Learnability with Prior Information 335
9.3.2 Some Simple Sufficient Conditions 337
9.3.3 Dispersability of Function Classes 341
9.3.4 Connections Between Dispersability and Learnability WPI 344
9.3.5 Distribution-Free Learning with Prior Information 348
9.4 Learning with Prior Information: Bounds on Learning Rates . 352

10. Applications to Neural Networks 365
10.1 What is a Neural Network? 366
10.2 Learning in Neural Networks 369
10.2.1 Problem Formulation 369
10.2.2 Reprise of Sample Complexity Estimates 372
10.2.3 Complexity-Theoretic Limits to Learnability 377
10.3 Estimates of VC-Dimensions of Families of Networks 381
10.3.1 Multi-Layer Perceptron Networks 382
10.3.2 A Network with Infinite VC-Dimension 388
10.3.3 Neural Networks as Verifiers of Formulas 390
10.3.4 Neural Networks with Piecewise-Polynomial Activation Functions 396
10.3.5 A General Approach 402
10.3.6 An Improved Bound 406
10.3.7 Networks with Pfaffian Activation Functions 410
10.3.8 Results Based on Order-Minimality 413
10.4 Structural Risk Minimization 415

11. Applications to Control Systems 421
11.1 Randomized Algorithms for Robustness Analysis 421
11.1.1 Introduction to Robust Control 421
11.1.2 Some NP-Hard Problems in Robust Control 424
11.1.3 Randomized Algorithms for Robustness Analysis 426
11.2 Randomized Algorithms for Robust Controller Synthesis: General Approach 429
11.2.1 Paradigm of Robust Controller Synthesis Problem 429
11.2.2 Various Types of "Near" Minima 432
11.2.3 A General Approach to Randomized Algorithms 435
11.2.4 Two Algorithms for Finding Probably Approximate Near Minima 436
11.3 VC-Dimension Estimates for Problems in Robust Controller Synthesis 438
11.3.1 A General Result 438
11.3.2 Robust Stabilization 438
11.3.3 Weighted H_∞-Norm Minimization 441
11.3.4 Weighted H_2-Norm Minimization 444
11.3.5 Sample Complexity Considerations 445
11.3.6 Robust Controller Design Using Randomized Algorithms: An Example 449
11.4 A Learning Theory Approach to System Identification 453
11.4.1 Problem Formulation 453
11.4.2 A General Result 455
11.4.3 Sufficient Conditions for the UCEM Property 458
11.4.4 Bounds on the P-Dimension 461

12. Some Open Problems 465