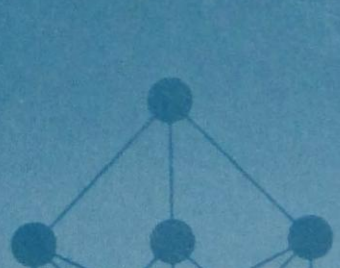
MACHINE LEARNING



TOM M. MITCHELL

McGRAW-HILL INTERNATIONAL EDITIONS
Computer Science Series

MACHINE LEARNING

Tom M. Mitchell
Carnegie Mellon University

The McGraw-Hill Companies, Inc.
New York St. Louis San Francisco Auckland Bogotá Caracas
Lisbon London Madrid Mexico City Milan Montreal New Delhi
San Juan Singapore Sydney Tokyo Toronto

CONTENTS

Preface xv

Acknowledgments xvi

1 Introduction 1

1.1 Well-Posed Learning Problems 2

1.2 Designing a Learning System 5

1.2.1 Choosing the Training Experience 5

1.2.2 Choosing the Target Function 7

1.2.3 Choosing a Representation for the Target Function 8

1.2.4 Choosing a Function Approximation Algorithm 9

1.2.5 The Final Design 11

1.3 Perspectives and Issues in Machine Learning 14

1.3.1 Issues in Machine Learning 15

1.4 How to Read This Book 16

1.5 Summary and Further Reading 17

Exercises 18

References 19

2 Concept Learning and the General-to-Specific Ordering 20

2.1 Introduction 20

2.2 A Concept Learning Task 21

2.2.1 Notation 22

2.2.2 The Inductive Learning Hypothesis 23

2.3 Concept Learning as Search 23

2.3.1 General-to-Specific Ordering of Hypotheses 24

2.4 FIND-S: Finding a Maximally Specific Hypothesis 26

2.5 Version Spaces and the CANDIDATE-ELIMINATION Algorithm 29

2.5.1 Representation 29

2.5.2 The LIST-THEN-ELIMINATE Algorithm 30

2.5.3 A More Compact Representation for Version Spaces 30

2.5.4 CANDIDATE-ELIMINATION Learning Algorithm 32
2.5.5 An Illustrative Example 33
2.6 Remarks on Version Spaces and CANDIDATE-ELIMINATION 37
2.6.1 Will the CANDIDATE-ELIMINATION Algorithm Converge to the Correct Hypothesis? 37
2.6.2 What Training Example Should the Learner Request Next? 37
2.6.3 How Can Partially Learned Concepts Be Used? 38
2.7 Inductive Bias 39
2.7.1 A Biased Hypothesis Space 40
2.7.2 An Unbiased Learner 40
2.7.3 The Futility of Bias-Free Learning 42
2.8 Summary and Further Reading 45
Exercises 47
References 50

3 Decision Tree Learning 52
3.1 Introduction 52
3.2 Decision Tree Representation 52
3.3 Appropriate Problems for Decision Tree Learning 54
3.4 The Basic Decision Tree Learning Algorithm 55
3.4.1 Which Attribute Is the Best Classifier? 55
3.4.2 An Illustrative Example 59
3.5 Hypothesis Space Search in Decision Tree Learning 60
3.6 Inductive Bias in Decision Tree Learning 63
3.6.1 Restriction Biases and Preference Biases 63
3.6.2 Why Prefer Short Hypotheses? 65
3.7 Issues in Decision Tree Learning 66
3.7.1 Avoiding Overfitting the Data 66
3.7.2 Incorporating Continuous-Valued Attributes 72
3.7.3 Alternative Measures for Selecting Attributes 73
3.7.4 Handling Training Examples with Missing Attribute Values 75
3.7.5 Handling Attributes with Differing Costs 75
3.8 Summary and Further Reading 76
Exercises 77
References 78

4 Artificial Neural Networks 81
4.1 Introduction 81
4.1.1 Biological Motivation 82
4.2 Neural Network Representations 82
4.3 Appropriate Problems for Neural Network Learning 83
4.4 Perceptrons 86
4.4.1 Representational Power of Perceptrons 86
4.4.2 The Perceptron Training Rule 88
4.4.3 Gradient Descent and the Delta Rule 89
4.4.4 Remarks 94

4.5 Multilayer Networks and the BACKPROPAGATION Algorithm 95
- 4.5.1 A Differentiable Threshold Unit 95
- 4.5.2 The BACKPROPAGATION Algorithm 97
- 4.5.3 Derivation of the BACKPROPAGATION Rule 101

4.6 Remarks on the BACKPROPAGATION Algorithm 104
- 4.6.1 Convergence and Local Minima 104
- 4.6.2 Representational Power of Feedforward Networks 105
- 4.6.3 Hypothesis Space Search and Inductive Bias 106
- 4.6.4 Hidden Layer Representations 106
- 4.6.5 Generalization, Overfitting, and Stopping Criterion 108

4.7 An Illustrative Example: Face Recognition 112
- 4.7.1 The Task 112
- 4.7.2 Design Choices 113
- 4.7.3 Learned Hidden Representations 116

4.8 Advanced Topics in Artificial Neural Networks 117
- 4.8.1 Alternative Error Functions 117
- 4.8.2 Alternative Error Minimization Procedures 119
- 4.8.3 Recurrent Networks 119
- 4.8.4 Dynamically Modifying Network Structure 121

4.9 Summary and Further Reading 122

Exercises 124

References 126

5 Evaluating Hypotheses 128

5.1 Motivation 128

5.2 Estimating Hypothesis Accuracy 129
- 5.2.1 Sample Error and True Error 130
- 5.2.2 Confidence Intervals for Discrete-Valued Hypotheses 131

5.3 Basics of Sampling Theory 132
- 5.3.1 Error Estimation and Estimating Binomial Proportions 133
- 5.3.2 The Binomial Distribution 135
- 5.3.3 Mean and Variance 136
- 5.3.4 Estimators, Bias, and Variance 137
- 5.3.5 Confidence Intervals 138
- 5.3.6 Two-Sided and One-Sided Bounds 141

5.4 A General Approach for Deriving Confidence Intervals 142
- 5.4.1 Central Limit Theorem 142

5.5 Difference in Error of Two Hypotheses 143
- 5.5.1 Hypothesis Testing 144

5.6 Comparing Learning Algorithms 145
- 5.6.1 Paired t Tests 148
- 5.6.2 Practical Considerations 149

5.7 Summary and Further Reading 150

Exercises 152

References 152

6 Bayesian Learning 154

6.1 Introduction 154

6.2 Bayes Theorem 156
- 6.2.1 An Example 157

6.3 Bayes Theorem and Concept Learning 158
6.3.1 Brute-Force Bayes Concept Learning 159
6.3.2 MAP Hypotheses and Consistent Learners 162
6.4 Maximum Likelihood and Least-Squared Error Hypotheses 164
6.5 Maximum Likelihood Hypotheses for Predicting Probabilities 167
6.5.1 Gradient Search to Maximize Likelihood in a Neural Net 170
6.6 Minimum Description Length Principle 171
6.7 Bayes Optimal Classifier 174
6.8 Gibbs Algorithm 176
6.9 Naive Bayes Classifier 177
6.9.1 An Illustrative Example 178
6.10 An Example: Learning to Classify Text 180
6.10.1 Experimental Results 182
6.11 Bayesian Belief Networks 184
6.11.1 Conditional Independence 185
6.11.2 Representation 186
6.11.3 Inference 187
6.11.4 Learning Bayesian Belief Networks 188
6.11.5 Gradient Ascent Training of Bayesian Networks 188
6.11.6 Learning the Structure of Bayesian Networks 190
6.12 The EM Algorithm 191
6.12.1 Estimating Means of k Gaussians 191
6.12.2 General Statement of EM Algorithm 194
6.12.3 Derivation of the k Means Algorithm 195
6.13 Summary and Further Reading 197
Exercises 198
References 199

7 Computational Learning Theory 201
7.1 Introduction 201
7.2 Probably Learning an Approximately Correct Hypothesis 203
7.2.1 The Problem Setting 203
7.2.2 Error of a Hypothesis 204
7.2.3 PAC Learnability 205
7.3 Sample Complexity for Finite Hypothesis Spaces 207
7.3.1 Agnostic Learning and Inconsistent Hypotheses 210
7.3.2 Conjunctions of Boolean Literals Are PAC-Learnable 211
7.3.3 PAC-Learnability of Other Concept Classes 212
7.4 Sample Complexity for Infinite Hypothesis Spaces 214
7.4.1 Shattering a Set of Instances 214
7.4.2 The Vapnik-Chervonenkis Dimension 215
7.4.3 Sample Complexity and the VC Dimension 217
7.4.4 VC Dimension for Neural Networks 218
7.5 The Mistake Bound Model of Learning 220
7.5.1 Mistake Bound for the FIND-S Algorithm 220
7.5.2 Mistake Bound for the HALVING Algorithm 221
7.5.3 Optimal Mistake Bounds 222
7.5.4 WEIGHTED-MAJORITY Algorithm 223

7.6 Summary and Further Reading 225
Exercises 227
References 229

8 Instance-Based Learning 230
8.1 Introduction 230
8.2 k-NEAREST NEIGHBOR LEARNING 231
8.2.1 Distance-Weighted NEAREST NEIGHBOR Algorithm 233
8.2.2 Remarks on k-NEAREST NEIGHBOR Algorithm 234
8.2.3 A Note on Terminology 236
8.3 Locally Weighted Regression 236
8.3.1 Locally Weighted Linear Regression 237
8.3.2 Remarks on Locally Weighted Regression 238
8.4 Radial Basis Functions 238
8.5 Case-Based Reasoning 240
8.6 Remarks on Lazy and Eager Learning 244
8.7 Summary and Further Reading 245
Exercises 247
References 247

9 Genetic Algorithms 249
9.1 Motivation 249
9.2 Genetic Algorithms 250
9.2.1 Representing Hypotheses 252
9.2.2 Genetic Operators 253
9.2.3 Fitness Function and Selection 255
9.3 An Illustrative Example 256
9.3.1 Extensions 258
9.4 Hypothesis Space Search 259
9.4.1 Population Evolution and the Schema Theorem 260
9.5 Genetic Programming 262
9.5.1 Representing Programs 262
9.5.2 Illustrative Example 263
9.5.3 Remarks on Genetic Programming 265
9.6 Models of Evolution and Learning 266
9.6.1 Lamarckian Evolution 266
9.6.2 Baldwin Effect 267
9.7 Parallelizing Genetic Algorithms 268
9.8 Summary and Further Reading 268
Exercises 270
References 270

10 Learning Sets of Rules 274
10.1 Introduction 274
10.2 Sequential Covering Algorithms 275
10.2.1 General to Specific Beam Search 277
10.2.2 Variations 279
10.3 Learning Rule Sets: Summary 280

10.4 Learning First-Order Rules 283
10.4.1 First-Order Horn Clauses 283
10.4.2 Terminology 284
10.5 Learning Sets of First-Order Rules: FOIL 285
10.5.1 Generating Candidate Specializations in FOIL 287
10.5.2 Guiding the Search in FOIL 288
10.5.3 Learning Recursive Rule Sets 290
10.5.4 Summary of FOIL 290
10.6 Induction as Inverted Deduction 291
10.7 Inverting Resolution 293
10.7.1 First-Order Resolution 296
10.7.2 Inverting Resolution: First-Order Case 297
10.7.3 Summary of Inverse Resolution 298
10.7.4 Generalization, θ-Subsumption, and Entailment 299
10.7.5 PROGOL 300
10.8 Summary and Further Reading 301
Exercises 303
References 304

11 Analytical Learning 307
11.1 Introduction 307
11.1.1 Inductive and Analytical Learning Problems 310
11.2 Learning with Perfect Domain Theories: PROLOG-EBG 312
11.2.1 An Illustrative Trace 313
11.3 Remarks on Explanation-Based Learning 319
11.3.1 Discovering New Features 320
11.3.2 Deductive Learning 321
11.3.3 Inductive Bias in Explanation-Based Learning 322
11.3.4 Knowledge Level Learning 323
11.4 Explanation-Based Learning of Search Control Knowledge 325
11.5 Summary and Further Reading 328
Exercises 330
References 331

12 Combining Inductive and Analytical Learning 334
12.1 Motivation 334
12.2 Inductive-Analytical Approaches to Learning 337
12.2.1 The Learning Problem 337
12.2.2 Hypothesis Space Search 339
12.3 Using Prior Knowledge to Initialize the Hypothesis 340
12.3.1 The KBANN Algorithm 340
12.3.2 An Illustrative Example 341
12.3.3 Remarks 344
12.4 Using Prior Knowledge to Alter the Search Objective 346
12.4.1 The TANGENTPROP Algorithm 347
12.4.2 An Illustrative Example 349
12.4.3 Remarks 350
12.4.4 The EBNN Algorithm 351
12.4.5 Remarks 355

12.5 Using Prior Knowledge to Augment Search Operators 357
12.5.1 The FOCL Algorithm 357
12.5.2 Remarks 360
12.6 State of the Art 361
12.7 Summary and Further Reading 362
Exercises 363
References 364

13 Reinforcement Learning 367
13.1 Introduction 367
13.2 The Learning Task 370
13.3 Q Learning 373
13.3.1 The Q Function 374
13.3.2 An Algorithm for Learning Q 374
13.3.3 An Illustrative Example 376
13.3.4 Convergence 377
13.3.5 Experimentation Strategies 379
13.3.6 Updating Sequence 379
13.4 Nondeterministic Rewards and Actions 381
13.5 Temporal Difference Learning 383
13.6 Generalizing from Examples 384
13.7 Relationship to Dynamic Programming 385
13.8 Summary and Further Reading 386
Exercises 388
References 388

Appendix Notation 391

Indexes
Author Index 394
Subject Index 400