Contents

Preface **ix**

Symbols and Notation **xiii**

1 Introduction **1**

2 Basics of Supervised Learning **5**
2.1 Introduction 5
2.2 Regression 7
2.2.1 Linear Models 7
2.2.2 Nonlinear Models 8
2.2.3 Error Measures 9
2.3 Classification 11
2.4 Time Series Prediction 13
2.5 Model Selection 14
2.6 Cross-validation 18
2.7 Further Reading 19

3 Basics of Symbolic Regression **21**
3.1 Example: Identification of a Polynomial 21
3.1.1 Data Collection and Preprocessing 22
3.1.2 Establishing a Baseline 22
3.1.3 Modeling Approach 23
3.1.4 Modeling Results 24
3.2 Example: Discovery of Laws of Physics from Data 26
3.3 Example: Approximation of the Gamma Function 29
3.4 Extending Symbolic Regression to Classification 32
3.4.1 Model Structures for Symbolic Classification 32
3.4.2 Evaluation of Symbolic Classification Models 33
3.5 Further Reading 34

4 Evolutionary Computation and Genetic Programming **35**
4.1 General Concepts 39
4.1.1 Genotype, Phenotype, and Semantics 39
4.1.2 Diversity and Evolvability 40
4.1.3 Buffering, Redundancy, and Neutrality 41
4.2 Population Initialization 42

4.2.1 Operators . . . 43
4.3 Fitness Calculation . . . 45
4.4 Parent Selection . . . 45
4.4.1 Operators . . . 45
4.4.2 Selection Pressure . . . 46
4.5 Bloat and Introns . . . 47
4.6 Crossover and Mutation . . . 49
4.7 Power of the Hypothesis Space . . . 50
4.8 GP Dynamics . . . 54
4.8.1 Fitness . . . 56
4.8.2 Variable Relevance . . . 56
4.8.3 Model Complexity . . . 58
4.8.4 Diversity . . . 59
4.9 Algorithmic Extensions . . . 64
4.9.1 Brood Selection and Offspring Selection . . . 64
4.9.2 Age-layered Population Structures . . . 69
4.9.3 Multi-objective GP . . . 70
4.9.4 Alternative Encodings: Linear and Graph GP . . . 73
4.9.5 Restricting Expressions: Syntax and Types . . . 74
4.9.6 Semantics-aware GP . . . 76
4.10 Conclusions . . . 77
4.11 Further Reading . . . 78

5 Model Validation, Inspection, Simplification, and Selection 81
5.1 Model Validation . . . 82
5.1.1 Visual Tools . . . 83
5.1.2 Explaining Models . . . 90
5.1.3 Model Interpretability . . . 99
5.2 Model Selection . . . 100
5.2.1 Criteria for Model Selection . . . 100
5.2.2 Hold-out Set for Validation . . . 101
5.2.3 Cross-validation . . . 101
5.2.4 Akaike's Information Criterion . . . 102
5.2.5 Bayesian Information Criterion . . . 104
5.2.6 Minimum Description Length Principle . . . 104
5.2.7 Comparison of Model Selection Criteria . . . 106
5.3 Model Simplification . . . 108
5.3.1 Nested Models . . . 109
5.3.2 Removal of Subexpressions . . . 111
5.4 Example: Boston Housing . . . 113
5.4.1 Data Preprocessing . . . 114
5.4.2 Model Generation and Selection for Median Values of Homes . . . 116
5.4.3 Model Generation and Selection for NO_X Concentrations 122
5.5 Conclusions . . . 125

5.6 Further Reading 125

6 Advanced Techniques **127**
6.1 Integration of Knowledge 127
6.1.1 Example Applications 128
6.1.2 Knowledge Integration Methods 128
6.1.3 Knowledge Integration via Customized Fitness Evaluation 129
6.1.4 Shape Constraints 131
6.1.5 Knowledge Integration via the Hypothesis Space . . . 133
6.2 Optimization of Coefficients 138
6.2.1 Linear Scaling 139
6.2.2 Nonlinear Optimization of Coefficients 141
6.3 Prediction Intervals 143
6.3.1 Prediction intervals for Linear Models 146
6.3.2 Approximate Prediction Intervals for Nonlinear Models 147
6.3.3 Bayesian Prediction Intervals 152
6.4 Modeling System Dynamics 153
6.4.1 Basics of Differential Equations 154
6.4.2 Finding Differential Equations with Symbolic Regression 157
6.5 Non-numeric Data 166
6.6 Non-evolutionary Symbolic Regression 168
6.6.1 Fast Function Extraction 168
6.6.2 Sparse Identification of Nonlinear Dynamics (SINDy) 169
6.6.3 Prioritized Grammar Enumeration 169
6.6.4 Exhaustive Symbolic Regression 169
6.6.5 Grammar-guided Exhaustive Equation Search 170
6.6.6 Equation Learner 171
6.6.7 Deep Symbolic Regression 171
6.6.8 AI Feynman 172

7 Examples and Applications **173**
7.1 Yacht Hydrodynamics 173
7.2 Industrial Chemical Processes 179
7.2.1 Correlation Analysis 180
7.2.2 Experimental Setup 181
7.2.3 Results for the Chemical Dataset 182
7.2.4 Results for the Tower Dataset 183
7.3 Interatomic Potentials 187
7.4 Friction 189
7.5 Lithium-ion Batteries 197
7.5.1 NASA PCoE Battery Datasets 197
7.5.2 First Version of the State-of-charge Model 199
7.5.3 Extended Version of the State-of-charge Model 203
7.5.4 Predicting the Discharge Voltage Curve 209

7.6 Biomedical Problems . . . 214
7.6.1 Identification of Virtual Tumor Markers . . . 216
7.7 Function Approximation . . . 219
7.8 Atmospheric CO_2 Concentration . . . 223
7.9 Flow Stress . . . 225
7.10 Dynamics of Simple Mechanical Systems . . . 229
7.10.1 Oscillator . . . 230
7.10.2 Pendulum . . . 233
7.10.3 Double Oscillator . . . 236
7.10.4 Double Pendulum . . . 239
7.11 Conclusions . . . 247

8 Conclusion 249
8.1 Unique Selling Points of Symbolic Regression . . . 249
8.2 Limitations and Caveats . . . 250

9 Appendix 253
9.1 Benchmarks . . . 253
9.2 Open-source Software for Genetic Programming . . . 254
9.2.1 HeuristicLab . . . 254
9.2.2 Operon . . . 255
9.2.3 PySR . . . 255
9.2.4 DEAP . . . 256
9.2.5 FEAT . . . 256
9.2.6 ECJ . . . 256
9.2.7 GPTIPS . . . 257
9.3 Commercial Software for Genetic Programming . . . 258
9.3.1 Evolved Analytics DataModeler . . . 258
9.3.2 Eureqa Formulize . . . 258

Bibliography 259

Index 285