

Wenmin Wang

Principles of Machine Learning

The Three Perspectives

 Springer

Contents

Part I Perspectives

1	Introduction	3
1.1	From “Queen” to “King”	3
1.1.1	Origins of Machine Learning	4
1.1.2	Definitions of Machine Learning	5
1.1.3	Machine Learning vs. Human Learning	6
1.2	Related Fields of Machine Learning	7
1.2.1	Artificial Intelligence	8
1.2.2	Deep Learning	9
1.2.3	Statistical Learning	10
1.2.4	Pattern Recognition	10
1.2.5	Data Mining	11
1.2.6	Others	12
1.3	History of Machine Learning	12
1.3.1	Neural Networks	12
1.3.2	Decision Trees	16
1.3.3	Boosting Methods	18
1.3.4	Support Vector Machines	19
1.3.5	Reinforcement Learning	21
1.4	Applications of Machine Learning	23
1.4.1	Computer Vision	23
1.4.2	Computer Speech	24
1.4.3	Natural Language Processing	25
1.4.4	Data Mining	25
1.4.5	Computer Games	26
1.4.6	Autonomous Technologies	30
1.4.7	Machine Learning in Medicine	32
1.4.8	Machine Learning in Biology	33
1.4.9	Others	34

1.5	The Too-Small World	34
	References	36
2	On Perspectives	41
2.1	Learning and Perspectives	41
2.1.1	Learning	41
2.1.2	Learning Theory	42
2.1.3	Machine Learning Theories	43
2.1.4	The Perspectives	44
2.2	Foundations of Machine Learning	44
2.2.1	Mathematics	45
2.2.2	Logic	49
2.2.3	Information Theory	50
2.2.4	Decision Theory	51
2.2.5	Neuroscience	52
2.3	Difficulties in Studying Machine Learning	53
2.3.1	Different Problems	53
2.3.2	Different Methods	54
2.3.3	Different Theories	54
2.4	Viewpoints on Machine Learning	54
2.4.1	What to Do	55
2.4.2	What Type	55
2.4.3	How to Do	55
2.5	Perspectives on Machine Learning	56
2.5.1	Learning Frameworks	56
2.5.2	Learning Paradigms	58
2.5.3	Learning Tasks	60
2.6	Relations Between the Perspectives	62
2.6.1	Logical Relationships	62
2.6.2	Hierarchical Relationship	64
	References	66

Part II Frameworks

3	Probabilistic Framework	69
3.1	Overview	69
3.1.1	Probability Theory	69
3.1.2	Probability Learning Theory	70
3.2	Basics of Probability Theory	71
3.2.1	Probability Space	72
3.2.2	Random Variable	73
3.2.3	Discrete and Continuous	74
3.2.4	Probability Relationship	75
3.2.5	Independence	77
3.2.6	Expectation	77
3.2.7	Variance and Standard Deviation	78

3.2.8	Covariance and Correlation	79
3.2.9	Convergence	80
3.3	Computational Learning Theory	82
3.3.1	Learnability Theory	83
3.3.2	PAC Learning	87
3.3.3	Occam Learning	89
3.4	Bayesian Models	91
3.4.1	Bayes' Theorem	91
3.4.2	Bayesian Learning	92
3.4.3	PAC-Bayesian Learning	93
3.5	Markov Models	94
3.5.1	Stochastic Process	95
3.5.2	Markov Property	98
3.5.3	Markov Process	99
3.5.4	Markov Chain	100
3.5.5	Hidden Markov Model	101
3.6	Probability Models	102
3.6.1	Discriminative Models	103
3.6.2	Generative Models	104
3.6.3	Comparative Analysis	105
3.7	Probabilistic Graphical Models	105
3.7.1	Probabilistic Graphs	106
3.7.2	Directed Graphical Models	107
3.7.3	Undirected Graphical Models	109
3.7.4	Factor Graphical Models	110
3.7.5	Comparative Analysis	111
3.8	Monte Carlo Methods	112
3.8.1	Origin of the Name	112
3.8.2	Monte Carlo Integration	113
3.8.3	Importance Sampling	115
3.8.4	Markov Chain Monte Carlo	117
3.8.5	Monte Carlo Tree Search	119
	References	122
4	Statistical Framework	125
4.1	Overview	125
4.1.1	Statistics	125
4.1.2	Statistical Learning	126
4.2	Descriptive Statistics	127
4.2.1	Central Tendency	128
4.2.2	Dispersion	130
4.3	Inferential Statistics	131
4.3.1	Parameter Estimation	132
4.3.2	Hypothesis Testing	136
4.4	Statistical Inference	137

4.4.1	Frequentist Inference	137
4.4.2	Bayesian Inference	139
4.4.3	Comparative Analysis	139
4.5	Statistical Learning Theory	140
4.5.1	Statistical Model	140
4.5.2	Statistical Learning Model	141
4.5.3	Growth Function	142
4.5.4	VC Dimension	144
4.5.5	Rademacher Complexity	146
4.6	Parametric and Nonparametric	150
4.6.1	Parametric Models	150
4.6.2	Nonparametric Models	151
4.6.3	Comparative Analysis	153
4.7	Kernel Methods	153
4.7.1	Kernel Trick	155
4.7.2	Hilbert Space	156
4.7.3	Kernel Functions	157
4.7.4	Positive-Definite Kernel	158
4.7.5	Reproducing Kernel	159
	References	161
5	Connectionist Framework	163
5.1	Overview	163
5.2	Connectionist Learning Theories	164
5.2.1	Hebbian Theory	164
5.2.2	Parallel Distributed Processing	165
5.2.3	Connectionist Models	166
5.2.4	Neural Network Theory	167
5.3	Units of Neural Networks	168
5.3.1	Biological Neuron	168
5.3.2	Artificial Neuron	169
5.3.3	Spiking Neuron	172
5.3.4	Long Short-Term Memory Unit	174
5.3.5	Gated Recurrent Unit	175
5.3.6	Capsule Unit	176
5.4	Structures of Neural Networks	178
5.4.1	Feedforward Neural Networks	178
5.4.2	Feedback Neural Networks	179
5.4.3	Symmetric Neural Networks	182
5.5	Optimizations of Neural Networks	182
5.5.1	Backpropagation	183
5.5.2	Evolutionary Methods	187
5.5.3	Weight-Agnostic Method	189
5.6	Agents of Neural Networks	190
5.6.1	Single- and Multiagent Networks	190

5.6.2	Case Study of Single-Agent Networks	190
5.6.3	Case Study of Multiagent Networks	192
	References	194
6	Symbolic Framework	197
6.1	Overview	197
6.1.1	Symbolism	198
6.1.2	Symbolic Learning Theory	198
6.1.3	Physical Symbol System	199
6.1.4	GOF AI	200
6.2	Formal Logic	201
6.2.1	Types of Formal Logic	201
6.2.2	Symbolic Logic	201
6.2.3	Mathematical Logic	204
6.3	Logic Learning	209
6.3.1	Logical Learning Theory	209
6.3.2	Logical Learning Modes	210
6.3.3	Logic Programming Methods	217
6.4	Rule Learning	222
6.4.1	Association Rule Learning	222
6.4.2	Association Rule Discovery	223
6.4.3	Association Rule Synthesis	224
6.5	Causal Learning	225
6.5.1	Causal Relations	225
6.5.2	Causal Inference	227
6.5.3	Causal Learning Approaches	229
6.6	Neuro-Symbolic Learning	232
6.6.1	Moravec's Paradox	232
6.6.2	Thinking Fast and Slow: System 1 and System 2	233
6.6.3	Neuro-Symbolic Approaches	235
	References	238
7	Behavioral Framework	241
7.1	Overview	241
7.1.1	Behaviorism Doctrine	241
7.1.2	Behavioral Psychology	243
7.2	Behavioral Learning Theory	243
7.2.1	Behavioral Learning	243
7.2.2	Respondent Conditioning	244
7.2.3	The Law of Effect	245
7.2.4	Operant Conditioning	246
7.2.5	Reinforcement Schedules	248
7.3	Behavioral Decision Theory	250
7.3.1	Decision Theory	251
7.3.2	Behavioral Decision	252
7.4	Behavioral Decision Process	252

7.4.1	St. Petersburg Paradox	253
7.4.2	Decision Process	254
7.4.3	Expected Utility	254
7.5	Bayesian Decision Models	256
7.5.1	Bayesian Decision Theory	256
7.5.2	Bayesian Decision Networks	258
7.6	Markov Decision Models	259
7.6.1	Markov Decision Process	260
7.6.2	Partially Observable Markov Decision Process	261
7.7	Behavioral Decision Types	261
7.7.1	Single-Stage Decision	262
7.7.2	Multistage Decision	262
7.7.3	Sequential Decision	263
	References	265

Part III Paradigms

8	Supervised Learning Paradigm	269
8.1	Definition	269
8.2	Working Principle	270
8.2.1	Formal Description	270
8.2.2	Illustrated Description	271
8.2.3	Operational Steps	272
8.3	Classic Tasks	272
8.4	Bias-Variance Problem	274
8.4.1	Bias and Variance	274
8.4.2	Bias-Variance Tradeoff	275
8.4.3	Bias-Variance Decomposition	276
8.5	Risk Minimization Principles	278
8.5.1	Expected Risk Minimization	279
8.5.2	Empirical Risk Minimization	280
8.5.3	Structural Risk Minimization	281
8.6	Variants of Supervised Learning	282
8.6.1	Weakly Supervised Learning	283
8.6.2	Semi-Supervised Learning	284
8.6.3	Label-Free Supervised Learning	284
8.7	No Free Lunch Theorems	286
8.7.1	Origin of the Theorems	286
8.7.2	Meaning of the Theorems	287
8.7.3	Two Sub-Theorems	287
8.7.4	Applied Scopes	289
	References	290

9	Unsupervised Learning Paradigm	291
9.1	Definition	291
9.2	Working Principle	292
9.2.1	Formal Description	292
9.2.2	Illustrated Description	293
9.3	Classic Tasks	294
9.4	Beyond the Classic	295
9.4.1	Importance	295
9.4.2	Generative Models	296
9.5	Energy Models	297
9.5.1	Basic Energy Model	298
9.5.2	Boltzmann Models	300
9.5.3	Related Works	301
9.6	Autoencoding Models	302
9.6.1	Autoencoders	302
9.6.2	Undercomplete Autoencoders	303
9.6.3	Sparse Autoencoders	304
9.6.4	Denoising Autoencoders	305
9.6.5	Variational Autoencoders	306
9.7	Generative Adversarial Models	310
9.7.1	Generative Adversarial Networks	310
9.7.2	Related Works	312
9.8	Normalizing Flow Models	312
9.8.1	Change of Variables	313
9.8.2	Normalizing Flow	314
9.8.3	Related Works	315
9.9	Autoregressive Models	315
9.9.1	k -Order Autoregressive Models	316
9.9.2	Deep Autoregressive Models	316
9.9.3	Related Works	317
9.10	Diffusion Models	318
9.10.1	Diffusion Probabilistic Models	318
9.10.2	Denoising Diffusion Processes	319
9.10.3	Training of Diffusion Models	320
9.10.4	Related Works	321
	References	323
10	Reinforcement Learning Paradigm	325
10.1	Overview	325
10.1.1	Ancient Reinforcement Learning	326
10.1.2	Neoteric Reinforcement Learning	326
10.1.3	Modern Reinforcement Learning	326
10.2	Definition	327
10.3	Related Elements	328
10.3.1	Policy	328

10.3.2	Return	329
10.3.3	Value Functions	331
10.3.4	Bellman Equations	332
10.3.5	Model-Based and Model-Free	335
10.3.6	On-Policy and Off-Policy	336
10.3.7	Hierarchical Relationships	337
10.4	Dynamic Programming	338
10.4.1	Value Iteration	338
10.4.2	Policy Iteration.....	339
10.4.3	Asynchronous Dynamic Programming	342
10.5	Monte Carlo Learning.....	343
10.5.1	Monte Carlo Prediction.....	344
10.5.2	Monte Carlo Control	345
10.6	Temporal Difference Learning.....	346
10.6.1	Temporal Difference Prediction.....	347
10.6.2	Temporal Difference Control	348
10.7	Eligibility Trace Mechanisms.....	349
10.7.1	Eligibility Trace.....	350
10.7.2	Forward View	350
10.7.3	Backward View	352
10.7.4	Comparative Analysis	352
10.8	Deep Reinforcement Learning.....	353
10.8.1	Deep Q -Network	353
10.8.2	Improved Algorithms and Networks.....	354
10.9	Comparison of Reinforcement Learning	356
	References.....	357
11	Other Learning Quasi-paradigm	359
11.1	Ensemble Learning.....	359
11.1.1	Definition	360
11.1.2	Combination Modes	360
11.1.3	Ensemble Methods	361
11.1.4	Averaging Scheme	363
11.1.5	Voting Scheme	364
11.2	Meta-Learning.....	366
11.2.1	Definition	366
11.2.2	Related Discourses	367
11.2.3	Related Terminologies.....	368
11.3	Transfer Learning	369
11.3.1	Definition	370
11.3.2	Working Principle	371
11.3.3	Categorization	373
11.4	Self-Supervised Learning	374
11.4.1	Definition	374
11.4.2	Related Discourses	375

11.4.3	Typical Methods	375
11.5	<i>n</i> -Shot Learning	376
11.5.1	Meaning of the Term	377
11.5.2	Related Subtasks	377
11.5.3	Working Principle	379
11.5.4	Case Study of One-Shot Learning	379
	References	381

Part IV Tasks

12	Classification Task	387
12.1	Problem and Definition	387
12.1.1	Classification Problems	387
12.1.2	Definition	388
12.2	Working Principle	388
12.2.1	Formal Description	388
12.2.2	Illustrated Description	389
12.2.3	Classification Functions	390
12.3	Basic Types	392
12.3.1	One-Class Classification	392
12.3.2	Two-Class Classification	392
12.3.3	Multi-class Classification	393
12.3.4	Multi-label Classification	394
12.3.5	Imbalanced Classification	395
12.4	Related Elements	395
12.4.1	Linear and Nonlinear	395
12.4.2	Hard and Soft	398
12.4.3	Lazy and Eager	400
12.5	Typical Algorithms	400
12.5.1	Logistic Regression	401
12.5.2	Naive Bayes	402
12.5.3	Adaptive Boosting	403
12.5.4	Support Vector Machine	405
12.5.5	Artificial Neural Networks	407
12.6	Loss Functions	407
12.6.1	0–1 Loss	408
12.6.2	Logistic Loss	409
12.6.3	Exponential Loss	409
12.6.4	Hinge Loss	409
12.6.5	Cross-Entropy Loss	410
12.7	Evaluation Metrics	411
12.7.1	Accuracy	413
12.7.2	Precision and Recall	414
12.7.3	Sensitivity and Specificity	414
12.7.4	F1 Score	415

12.7.5	P-R Curve	415
12.7.6	ROC Curve	416
12.7.7	Matthew's Correlation Coefficient	417
12.8	Application Fields	418
	References	419
13	Regression Task	421
13.1	Problem and Definition	421
13.1.1	Regression Problems	421
13.1.2	Definition	422
13.1.3	Regression Model	423
13.1.4	Underlying Assumptions	423
13.2	Working Principle	424
13.2.1	Formal Description	424
13.2.2	Illustrated Description	425
13.3	Related Elements	426
13.3.1	Unique and Multiple	426
13.3.2	Linear and Nonlinear	427
13.3.3	Parametric and Nonparametric	428
13.3.4	Interpolation and Extrapolation	429
13.4	Typical Algorithms	430
13.4.1	Multiple Linear Regression	431
13.4.2	Polynomial Regression	433
13.4.3	Ridge Regression	434
13.4.4	Lasso Regression	436
13.4.5	Bayesian Regression	438
13.5	Loss Functions	441
13.5.1	Mean Square Error	441
13.5.2	Mean Absolute Error	443
13.5.3	Mean Bias Error	443
13.5.4	Relative Absolute Error	444
13.5.5	Relative Squared Error	445
13.5.6	Huber Loss	445
13.5.7	Log-Cosh Loss	446
13.5.8	Coefficient of Determination	446
13.6	Application Fields	447
14	Clustering Task	449
14.1	Problem and Definition	449
14.1.1	Clustering Problems	449
14.1.2	Definition	450
14.2	Working Principle	450
14.2.1	Formal Description	450
14.2.2	Illustrated Description	451
14.3	Related Elements	452
14.3.1	Hard and Soft	452

14.3.2	Linear and Nonlinear	452
14.4	Classical Methods	452
14.4.1	Hierarchical-Based Clustering	453
14.4.2	Centroid-Based Clustering	455
14.4.3	Distribution-Based Clustering	455
14.4.4	Density-Based Clustering	456
14.4.5	Grid-Based Clustering	457
14.5	Neo-classical Methods	458
14.5.1	Quantum Clustering	458
14.5.2	Spectral Clustering	459
14.5.3	Swarm Intelligence Clustering	460
14.5.4	Deep Clustering	461
14.6	Typical Algorithms	462
14.6.1	k -Means	462
14.6.2	Gaussian Mixture Clustering	464
14.6.3	DBSCAN	465
14.6.4	Density Peak Clustering	467
14.7	Evaluation Metrics	468
14.7.1	Adjusted Rand Index	469
14.7.2	Adjusted Mutual Information	470
14.7.3	Fowlkes-Mallows Index	472
14.7.4	V-Measure	474
14.7.5	Calinski-Harabasz Index	475
14.7.6	Davies-Bouldin Index	476
14.7.7	Silhouette Coefficient	477
14.8	Application Fields	477
	References	478
15	Dimensionality Reduction Task	481
15.1	Problem and Definition	481
15.1.1	Dimensionality Reduction Problems	481
15.1.2	Definition	482
15.2	Working Principle	483
15.2.1	Formal Description	483
15.2.2	Illustrated Description	484
15.3	Related Elements	484
15.3.1	Feature Selection and Extraction	484
15.3.2	Linear and Nonlinear	485
15.3.3	Shallow and Deep	485
15.3.4	Hierarchical Relationship	485
15.4	Linear Dimensionality Reduction	486
15.4.1	Linear Subspace Learning	486
15.4.2	Multilinear Subspace Learning	487
15.5	Nonlinear Dimensionality Reduction	488
15.5.1	Distance Matrix Method	488

15.5.2	Kernel Methods	489
15.5.3	Manifold Learning	490
15.5.4	Neighborhood Embedding	491
15.6	Deep Dimensionality Reduction	492
15.6.1	Deep Autoencoder	492
15.6.2	Variational Autoencoder	493
15.7	Typical Algorithms	493
15.7.1	Principal Component Analysis	493
15.7.2	Multidimensional Scaling	495
15.7.3	Isometric Feature Mapping	498
15.7.4	Local Linear Embedding	499
15.7.5	Stochastic Neighbor Embedding	499
15.7.6	t -Distributed Stochastic Neighbor Embedding	501
15.8	Application Fields	504
	References	505
A	Common Terms	507
A.1	Algorithm, Model, and Hypothesis	507
A.1.1	Algorithm	507
A.1.2	Model	507
A.1.3	Hypothesis	508
A.2	Data	508
A.2.1	Datasets	508
A.2.2	Training, Validation, and Test Dataset	508
A.2.3	Labeling	509
A.2.4	Ground-Truth	509
A.2.5	Dataset vs. Database	510
A.3	Features	510
A.3.1	Hand-Crafted Features	510
A.3.2	Learned Features	511
A.4	Dimensionality	511
A.4.1	Mathematical Space	512
A.4.2	Data Space	512
A.5	Normalization and Regularization	512
A.5.1	Normalization	512
A.5.2	Regularization	513
A.6	Baseline	514
	References	515
	Glossary	517