

Machine Learning

Zhi-Hua Zhou

Machine Learning

Zhi-Hua Zhou
Nanjing University
Nanjing, Jiangsu, China

Translated by
Shaowu Liu
University of Technology Sydney
Ultimo, NSW, Australia

ISBN 978-981-15-1966-6 ISBN 978-981-15-1967-3 (eBook)
<https://doi.org/10.1007/978-981-15-1967-3>

Translation from the language edition: *Machine Learning* by Zhi-Hua Zhou, and Shaowu Liu, © Tsinghua University Press 2016. Published by Tsinghua University Press. All Rights Reserved.

© Springer Nature Singapore Pte Ltd. 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

This is an introductory-level machine learning textbook. To make the content accessible to a wider readership, the author has tried to reduce the use of mathematics. However, to gain a decent understanding of machine learning, basic knowledge of probability, statistics, algebra, optimization, and logic seems unavoidable. Therefore, this book is more appropriate for advanced undergraduate or graduate students in science and engineering, as well as practitioners and researchers with equivalent background knowledge.

The book has 16 chapters that can be roughly divided into three parts. The first part includes Chapters 1–3, which introduces the basics of machine learning. The second part includes Chapters 4–10, which presents some classic and popular machine learning methods. The third part includes Chapters 11–16, which covers advanced topics. As a textbook, Chapters 1–9 and 10 can be taught in one semester at the undergraduate level, while the whole book could be used for the graduate level.

This introductory textbook aims to cover the core topics of machine learning in one semester, and hence is unable to provide detailed discussions on many important frontier research works. The author believes that, for readers new to this field, it is more important to have a broad view than drill down into the very details. Hence, in-depth discussions are left to advanced courses. However, readers who wish to explore the topics of interest are encouraged to follow the further reading section at the end of each chapter.

The book was originally published in Chinese and had a wide readership in the Chinese community. The author would like to thank Dr. Shaowu Liu for his great effort of translating the book into English and thank Springer for the publication.

Zhi-Hua Zhou
Nanjing, China

Contents

1	Introduction	1
1.1	Introduction	2
1.2	Terminology	2
1.3	Hypothesis Space	5
1.4	Inductive Bias	7
1.5	Brief History	11
1.6	Application Status	15
1.7	Further Reading	19
	Exercises	21
	Break Time	22
	References	23
2	Model Selection and Evaluation	25
2.1	Empirical Error and Overfitting	26
2.2	Evaluation Methods	27
2.2.1	Hold-Out	28
2.2.2	Cross-Validation	29
2.2.3	Bootstrapping	30
2.2.4	Parameter Tuning and Final Model	31
2.3	Performance Measure	32
2.3.1	Error Rate and Accuracy	32
2.3.2	Precision, Recall, and $F1$	33
2.3.3	ROC and AUC	36
2.3.4	Cost-Sensitive Error Rate and Cost Curve	39
2.4	Comparison Test	41
2.4.1	Hypothesis Testing	41
2.4.2	Cross-Validated t -Test	44
2.4.3	McNemar's Test	45
2.4.4	Friedman Test and Nemenyi Post-hoc Test	46
2.5	Bias and Variance	49
2.6	Further Reading	52
	Exercises	53
	Break Time	54
	References	55
3	Linear Models	57
3.1	Basic Form	58
3.2	Linear Regression	58
3.3	Logistic Regression	62
3.4	Linear Discriminant Analysis	65
3.5	Multiclass Classification	68
3.6	Class Imbalance Problem	71
3.7	Further Reading	73

	Exercises	75
	Break Time	76
	References	77
4	Decision Trees	79
4.1	Basic Process	80
4.2	Split Selection	81
4.2.1	Information Gain	82
4.2.2	Gain Ratio	84
4.2.3	Gini Index	86
4.3	Pruning	86
4.3.1	Pre-pruning	88
4.3.2	Post-pruning	89
4.4	Continuous and Missing Values	90
4.4.1	Handling Continuous Values	90
4.4.2	Handling Missing Values	92
4.5	Multivariate Decision Trees	95
4.6	Further Reading	98
	Exercises	100
	Break Time	101
	References	102
5	Neural Networks	103
5.1	Neuron Model	104
5.2	Perceptron and Multi-layer Network	105
5.3	Error Backpropagation Algorithm	108
5.4	Global Minimum and Local Minimum	113
5.5	Other Common Neural Networks	115
5.5.1	RBF Network	115
5.5.2	ART Network	116
5.5.3	SOM Networks	117
5.5.4	Cascade-Correlation Network	118
5.5.5	Elman Network	119
5.5.6	Boltzmann Machine	119
5.6	Deep Learning	121
5.7	Further Reading	123
	Exercises	125
	Break Time	126
	References	127
6	Support Vector Machine	129
6.1	Margin and Support Vector	130
6.2	Dual Problem	132
6.3	Kernel Function	134
6.4	Soft Margin and Regularization	138
6.5	Support Vector Regression	142
6.6	Kernel Methods	145
6.7	Further Reading	148

	Exercises	150
	Break Time	151
	References	152
7	Bayes Classifiers	155
7.1	Bayesian Decision Theory	156
7.2	Maximum Likelihood Estimation	158
7.3	Naïve Bayes Classifier	159
7.4	Semi-Naïve Bayes Classifier	163
7.5	Bayesian Network	166
7.5.1	Network Structure	166
7.5.2	Learning	168
7.5.3	Inference	170
7.6	EM Algorithm	172
7.7	Further Reading	173
	Exercises	176
	Break Time	177
	References	178
8	Ensemble Learning	181
8.1	Individual and Ensemble	182
8.2	Boosting	184
8.3	Bagging and Random Forest	189
8.3.1	Bagging	190
8.3.2	Random Forest	191
8.4	Combination Strategies	193
8.4.1	Averaging	194
8.4.2	Voting	195
8.4.3	Combining by Learning	196
8.5	Diversity	198
8.5.1	Error-Ambiguity Decomposition	198
8.5.2	Diversity Measures	200
8.5.3	Diversity Generation	201
8.6	Further Reading	203
	Exercises	206
	Break Time	208
	References	209
9	Clustering	211
9.1	Clustering Problem	212
9.2	Performance Measure	213
9.3	Distance Calculation	215
9.4	Prototype Clustering	217
9.4.1	<i>k</i> -Means Clustering	217
9.4.2	Learning Vector Quantization	219
9.4.3	Mixture-of-Gaussian Clustering	222
9.5	Density Clustering	227

9.6	Hierarchical Clustering	231
9.7	Further Reading	234
	Exercises	236
	Break Time	238
	References	239
10	Dimensionality Reduction and Metric Learning	241
10.1	k-Nearest Neighbor Learning	242
10.2	Low-Dimensional Embedding	243
10.3	Principal Component Analysis	247
10.4	Kernelized PCA	250
10.5	Manifold Learning	252
10.5.1	Isometric Mapping	252
10.5.2	Locally Linear Embedding	254
10.6	Metric Learning	256
10.7	Further Reading	259
	Exercises	261
	Break Time	262
	References	263
11	Feature Selection and Sparse Learning	265
11.1	Subset Search and Evaluation	266
11.2	Filter Methods	268
11.3	Wrapper Methods	270
11.4	Embedded Methods and L_1 Regularization	271
11.5	Sparse Representation and Dictionary Learning	274
11.6	Compressed Sensing	276
11.7	Further Reading	280
	Exercises	282
	Break Time	283
	References	284
12	Computational Learning Theory	287
12.1	Basic Knowledge	288
12.2	PAC Learning	289
12.3	Finite Hypothesis Space	292
12.3.1	Separable Case	292
12.3.2	Non-separable Case	293
12.4	VC Dimension	295
12.5	Rademacher Complexity	300
12.6	Stability	306
12.7	Further Reading	309
	Exercises	311
	Break Time	312
	References	313

13	Semi-Supervised Learning	315
13.1	Unlabeled Samples	316
13.2	Generative Methods	319
13.3	Semi-Supervised SVM	321
13.4	Graph-Based Semi-Supervised Learning	324
13.5	Disagreement-Based Methods	328
13.6	Semi-Supervised Clustering	331
13.7	Further Reading	334
	Exercises	337
	Break Time	339
	References	340
14	Probabilistic Graphical Models	343
14.1	Hidden Markov Model	344
14.2	Markov Random Field	347
14.3	Conditional Random Field	351
14.4	Learning and Inference	353
14.4.1	Variable Elimination	354
14.4.2	Belief Propagation	356
14.5	Approximate Inference	357
14.5.1	MCMC Sampling	357
14.5.2	Variational Inference	360
14.6	Topic Model	363
14.7	Further Reading	366
	Exercises	368
	Break Time	369
	References	370
15	Rule Learning	373
15.1	Basic Concepts	374
15.2	Sequential Covering	376
15.3	Pruning Optimization	379
15.4	First-Order Rule Learning	381
15.5	Inductive Logic Programming	385
15.5.1	Least General Generalization	386
15.5.2	Inverse Resolution	388
15.6	Further Reading	391
	Exercises	394
	Break Time	396
	References	397
16	Reinforcement Learning	399
16.1	Task and Reward	400
16.2	K -Armed Bandit	402
16.2.1	Exploration Versus Exploitation	402
16.2.2	ϵ -Greedy	404
16.2.3	Softmax	405

16.3	Model-Based Learning	407
16.3.1	Policy Evaluation	407
16.3.2	Policy Improvement	410
16.3.3	Policy Iteration and Value Iteration	411
16.4	Model-Free Learning	413
16.4.1	Monte Carlo Reinforcement Learning	413
16.4.2	Temporal Difference Learning	418
16.5	Value Function Approximation	419
16.6	Imitation Learning	421
16.6.1	Direct Imitation Learning	422
16.6.2	Inverse Reinforcement Learning	423
16.7	Further Reading	424
	Exercises	426
	Break Time	428
	References	429
	Appendix A: Matrix	432
	Appendix B: Optimization	437
	Appendix C: Probability Distributions	445
	Index	453

Symbols

x	Scalar	$\ \cdot \ _p$	L_p norm; L_2 norm when p is absent
$\mathbf{x}$	Vector		
$\mathbf{x}$	Variable set	$P(\cdot), P(\cdot \cdot)$	Probability mass function, conditional probability mass function
$\mathbf{A}$	Matrix		
$\mathbf{I}$	Identity matrix		
$\mathcal{X}$	Sample space or state space	$p(\cdot), p(\cdot \cdot)$	Probability density function, conditional probability density function
$\mathcal{D}$	Probability distribution		
D	Data set		
$\mathcal{H}$	Hypothesis space	$\mathbb{E}_{\cdot \sim \mathcal{D}}[f(\cdot)]$	Expectation of function $f(\cdot)$ with respect to $\cdot$ over distribution $\mathcal{D}$; $\mathcal{D}$ and/or $\cdot$ are omitted when context is clear
H	Hypothesis set		
$\mathcal{L}$	Learning algorithm		
$(\cdot, \cdot, \cdot)$	Row vector	$\sup(\cdot)$	Supremum
$(\cdot; \cdot; \cdot)$	Column vector	$\mathbb{I}(\cdot)$	Indicator function, returns 1 if $\cdot$ is true, and 0 if $\cdot$ is false
$(\cdot)^T$	Transpose of vector or matrix	$\text{sign}(\cdot)$	Sign function, returns -1 if $\cdot < 0$, 0 if $\cdot = 0$, and 1 if $\cdot > 0$
$\{\cdots\}$	Set		
$ \{\cdots\} $	Number of elements in set $\{\cdots\}$		