

INTRODUCTION TO

Machine Learning

fourth edition

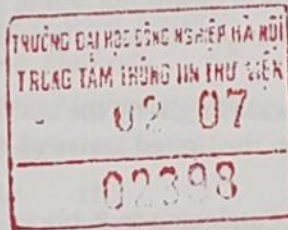
ETHEM ALPAYDIN

Introduction
to
Machine
Learning

Fourth
Edition

Ethem Alpaydm

The MIT Press
Cambridge, Massachusetts
London, England



Contents

Preface xix

Notations xxiii

1 Introduction 1

- 1.1 What Is Machine Learning? 1
- 1.2 Examples of Machine Learning Applications 4
 - 1.2.1 Association Rules 4
 - 1.2.2 Classification 4
 - 1.2.3 Regression 9
 - 1.2.4 Unsupervised Learning 11
 - 1.2.5 Reinforcement Learning 12
- 1.3 History 13
- 1.4 Related Topics 15
 - 1.4.1 High-Performance Computing 15
 - 1.4.2 Data Privacy and Security 16
 - 1.4.3 Model Interpretability and Trust 17
 - 1.4.4 Data Science 18
- 1.5 Exercises 18
- 1.6 References 20

2 Supervised Learning 23

- 2.1 Learning a Class from Examples 23
- 2.2 Vapnik-Chervonenkis Dimension 29
- 2.3 Probably Approximately Correct Learning 31
- 2.4 Noise 32
- 2.5 Learning Multiple Classes 34

2.6	Regression	36
2.7	Model Selection and Generalization	39
2.8	Dimensions of a Supervised Machine Learning Algorithm	43
2.9	Notes	44
2.10	Exercises	45
2.11	References	49
3	<i>Bayesian Decision Theory</i>	51
3.1	Introduction	51
3.2	Classification	53
3.3	Losses and Risks	55
3.4	Discriminant Functions	57
3.5	Association Rules	58
3.6	Notes	61
3.7	Exercises	62
3.8	References	66
4	<i>Parametric Methods</i>	67
4.1	Introduction	67
4.2	Maximum Likelihood Estimation	68
4.2.1	Bernoulli Density	69
4.2.2	Multinomial Density	70
4.2.3	Gaussian (Normal) Density	70
4.3	Evaluating an Estimator: Bias and Variance	71
4.4	The Bayes' Estimator	72
4.5	Parametric Classification	75
4.6	Regression	79
4.7	Tuning Model Complexity: Bias/Variance Dilemma	82
4.8	Model Selection Procedures	85
4.9	Notes	89
4.10	Exercises	90
4.11	References	93
5	<i>Multivariate Methods</i>	95
5.1	Multivariate Data	95
5.2	Parameter Estimation	96
5.3	Estimation of Missing Values	97
5.4	Multivariate Normal Distribution	98
5.5	Multivariate Classification	102
5.6	Tuning Complexity	108

5.7	Discrete Features	110
5.8	Multivariate Regression	111
5.9	Notes	113
5.10	Exercises	114
5.11	References	116
6	<i>Dimensionality Reduction</i>	117
6.1	Introduction	117
6.2	Subset Selection	118
6.3	Principal Component Analysis	122
6.4	Feature Embedding	129
6.5	Factor Analysis	132
6.6	Singular Value Decomposition and Matrix Factorization	137
6.7	Multidimensional Scaling	138
6.8	Linear Discriminant Analysis	142
6.9	Canonical Correlation Analysis	147
6.10	Isomap	150
6.11	Locally Linear Embedding	152
6.12	Laplacian Eigenmaps	154
6.13	t -Distributed Stochastic Neighbor Embedding	157
6.14	Notes	159
6.15	Exercises	161
6.16	References	162
7	<i>Clustering</i>	165
7.1	Introduction	165
7.2	Mixture Densities	166
7.3	k -Means Clustering	167
7.4	Expectation-Maximization Algorithm	171
7.5	Mixtures of Latent Variable Models	176
7.6	Supervised Learning after Clustering	177
7.7	Spectral Clustering	179
7.8	Hierarchical Clustering	180
7.9	Choosing the Number of Clusters	183
7.10	Notes	183
7.11	Exercises	184
7.12	References	186
8	<i>Nonparametric Methods</i>	189
8.1	Introduction	189

8.2	Nonparametric Density Estimation	190
8.2.1	Histogram Estimator	191
8.2.2	Kernel Estimator	192
8.2.3	k -Nearest Neighbor Estimator	194
8.3	Generalization to Multivariate Data	196
8.4	Nonparametric Classification	197
8.5	Condensed Nearest Neighbor	198
8.6	Distance-Based Classification	200
8.7	Outlier Detection	203
8.8	Nonparametric Regression: Smoothing Models	205
8.8.1	Running Mean Smoother	205
8.8.2	Kernel Smoother	207
8.8.3	Running Line Smoother	208
8.9	How to Choose the Smoothing Hyperparameter	208
8.10	Notes	209
8.11	Exercises	212
8.12	References	214
9	Decision Trees	217
9.1	Introduction	217
9.2	Univariate Trees	219
9.2.1	Classification Trees	220
9.2.2	Regression Trees	224
9.3	Pruning	226
9.4	Rule Extraction from Trees	229
9.5	Learning Rules from Data	230
9.6	Multivariate Trees	234
9.7	Notes	236
9.8	Exercises	239
9.9	References	241
10	Linear Discrimination	243
10.1	Introduction	243
10.2	Generalizing the Linear Model	245
10.3	Geometry of the Linear Discriminant	246
10.3.1	Two Classes	246
10.3.2	Multiple Classes	248
10.4	Pairwise Separation	250
10.5	Parametric Discrimination Revisited	251

10.6	Gradient Descent	252
10.7	Logistic Discrimination	254
10.7.1	Two Classes	254
10.7.2	Multiple Classes	257
10.7.3	Multiple Labels	263
10.8	Learning to Rank	264
10.9	Notes	265
10.10	Exercises	267
10.11	References	269
11	Multilayer Perceptrons	271
11.1	Introduction	271
11.1.1	Understanding the Brain	272
11.1.2	Neural Networks as a Paradigm for Parallel Processing	273
11.2	The Perceptron	275
11.3	Training a Perceptron	278
11.4	Learning Boolean Functions	282
11.5	Multilayer Perceptrons	283
11.6	MLP as a Universal Approximator	286
11.7	Backpropagation Algorithm	288
11.7.1	Nonlinear Regression	288
11.7.2	Two-Class Discrimination	291
11.7.3	Multiclass Discrimination	292
11.7.4	Multilabel Discrimination	294
11.8	Overtraining	295
11.9	Learning Hidden Representations	296
11.10	Autoencoders	301
11.11	Word2vec Architecture	303
11.12	Notes	307
11.13	Exercises	309
11.14	References	310
12	Deep Learning	313
12.1	Introduction	313
12.2	How to Train Multiple Hidden Layers	317
12.2.1	Rectified Linear Unit	317
12.2.2	Initialization	317

12.2.3	Generalizing Backpropagation to Multiple Hidden Layers	318	
12.3	Improving Training Convergence	321	
12.3.1	Momentum	321	
12.3.2	Adaptive Learning Factor	322	
12.3.3	Batch Normalization	323	
12.4	Regularization	325	
12.4.1	Hints	325	
12.4.2	Weight Decay	327	
12.4.3	Dropout	330	
12.5	Convolutional Layers	331	
12.5.1	The Idea	331	
12.5.2	Formalization	333	
12.5.3	Examples: LeNet-5 and AlexNet	337	
12.5.4	Extensions	338	
12.5.5	Multimodal Deep Networks	340	
12.6	Tuning the Network Structure	340	
12.6.1	Structure and Hyperparameter Search	340	
12.6.2	Skip Connections	342	
12.6.3	Gating Units	343	
12.7	Learning Sequences	344	
12.7.1	Example Tasks	344	
12.7.2	Time-Delay Neural Networks	345	
12.7.3	Recurrent Networks	345	
12.7.4	Long Short-Term Memory Unit	348	
12.7.5	Gated Recurrent Unit	349	
12.8	Generative Adversarial Network	350	
12.9	Notes	353	
12.10	Exercises	354	
12.11	References	356	
13	Local Models	361	
13.1	Introduction	361	
13.2	Competitive Learning	362	
13.2.1	Online k -Means	362	
13.2.2	Adaptive Resonance Theory	367	
13.2.3	Self-Organizing Maps	368	
13.3	Radial Basis Functions	370	
13.4	Incorporating Rule-Based Knowledge	376	

13.5	Normalized Basis Functions	377
13.6	Competitive Basis Functions	379
13.7	Learning Vector Quantization	382
13.8	The Mixture of Experts	382
13.8.1	Cooperative Experts	385
13.8.2	Competitive Experts	386
13.9	Hierarchical Mixture of Experts and Soft Decision Trees	386
13.10	Notes	388
13.11	Exercises	389
13.12	References	392
14	Kernel Machines	395
14.1	Introduction	395
14.2	Optimal Separating Hyperplane	397
14.3	The Nonseparable Case: Soft Margin Hyperplane	401
14.4	ν -SVM	404
14.5	Kernel Trick	405
14.6	Vectorial Kernels	407
14.7	Defining Kernels	410
14.8	Multiple Kernel Learning	411
14.9	Multiclass Kernel Machines	413
14.10	Kernel Machines for Regression	414
14.11	Kernel Machines for Ranking	419
14.12	One-Class Kernel Machines	420
14.13	Large Margin Nearest Neighbor Classifier	423
14.14	Kernel Dimensionality Reduction	425
14.15	Notes	426
14.16	Exercises	428
14.17	References	429
15	Graphical Models	433
15.1	Introduction	433
15.2	Canonical Cases for Conditional Independence	435
15.3	Generative Models	442
15.4	d-Separation	445
15.5	Belief Propagation	445
15.5.1	Chains	446
15.5.2	Trees	448
15.5.3	Polytrees	450

15.5.4	Junction Trees	452	
15.6	Undirected Graphs: Markov Random Fields	453	
15.7	Learning the Structure of a Graphical Model	456	
15.8	Influence Diagrams	457	
15.9	Notes	458	
15.10	Exercises	459	
15.11	References	461	
16	<i>Hidden Markov Models</i>	463	
16.1	Introduction	463	
16.2	Discrete Markov Processes	464	
16.3	Hidden Markov Models	467	
16.4	Three Basic Problems of HMMs	469	
16.5	Evaluation Problem	469	
16.6	Finding the State Sequence	473	
16.7	Learning Model Parameters	475	
16.8	Continuous Observations	478	
16.9	The HMM as a Graphical Model	479	
16.10	Model Selection in HMMs	482	
16.11	Notes	484	
16.12	Exercises	486	
16.13	References	489	
17	<i>Bayesian Estimation</i>	491	
17.1	Introduction	491	
17.2	Bayesian Estimation of the Parameters of a Discrete Distribution	495	
17.2.1	$K > 2$ States: Dirichlet Distribution	495	
17.2.2	$K = 2$ States: Beta Distribution	496	
17.3	Bayesian Estimation of the Parameters of a Gaussian Distribution	497	
17.3.1	Univariate Case: Unknown Mean, Known Variance	497	
17.3.2	Univariate Case: Unknown Mean, Unknown Variance	499	
17.3.3	Multivariate Case: Unknown Mean, Unknown Covariance	501	
17.4	Bayesian Estimation of the Parameters of a Function	502	
17.4.1	Regression	502	

17.4.2	Regression with Prior on Noise Precision	506
17.4.3	The Use of Basis/Kernel Functions	507
17.4.4	Bayesian Classification	509
17.5	Choosing a Prior	512
17.6	Bayesian Model Comparison	513
17.7	Bayesian Estimation of a Mixture Model	516
17.8	Nonparametric Bayesian Modeling	519
17.9	Gaussian Processes	520
17.10	Dirichlet Processes and Chinese Restaurants	524
17.11	Latent Dirichlet Allocation	526
17.12	Beta Processes and Indian Buffets	528
17.13	Notes	529
17.14	Exercises	530
17.15	References	531
18	Combining Multiple Learners	533
18.1	Rationale	533
18.2	Generating Diverse Learners	534
18.3	Model Combination Schemes	537
18.4	Voting	538
18.5	Error-Correcting Output Codes	542
18.6	Bagging	544
18.7	Boosting	545
18.8	The Mixture of Experts Revisited	548
18.9	Stacked Generalization	550
18.10	Fine-Tuning an Ensemble	551
18.10.1	Choosing a Subset of the Ensemble	552
18.10.2	Constructing Metalearners	552
18.11	Cascading	553
18.12	Notes	555
18.13	Exercises	557
18.14	References	559
19	Reinforcement Learning	563
19.1	Introduction	563
19.2	Single State Case: K -Armed Bandit	565
19.3	Elements of Reinforcement Learning	566
19.4	Model-Based Learning	569
19.4.1	Value Iteration	569

19.4.2	Policy Iteration	570	
19.5	Temporal Difference Learning	571	
19.5.1	Exploration Strategies	571	
19.5.2	Deterministic Rewards and Actions	572	
19.5.3	Nondeterministic Rewards and Actions	573	
19.5.4	Eligibility Traces	576	
19.6	Generalization	577	
19.7	Partially Observable States	580	
19.7.1	The Setting	580	
19.7.2	Example: The Tiger Problem	582	
19.8	Deep Q Learning	587	
19.9	Policy Gradients	588	
19.10	Learning to Play Backgammon and Go	591	
19.11	Notes	592	
19.12	Exercises	593	
19.13	References	595	
20	<i>Design and Analysis of Machine Learning Experiments</i>	597	
20.1	Introduction	597	
20.2	Factors, Response, and Strategy of Experimentation	600	
20.3	Response Surface Design	603	
20.4	Randomization, Replication, and Blocking	604	
20.5	Guidelines for Machine Learning Experiments	605	
20.6	Cross-Validation and Resampling Methods	608	
20.6.1	K-Fold Cross-Validation	609	
20.6.2	5 × 2 Cross-Validation	610	
20.6.3	Bootstrapping	611	
20.7	Measuring Classifier Performance	611	
20.8	Interval Estimation	614	
20.9	Hypothesis Testing	618	
20.10	Assessing a Classification Algorithm's Performance	620	
20.10.1	Binomial Test	621	
20.10.2	Approximate Normal Test	622	
20.10.3	t Test	622	
20.11	Comparing Two Classification Algorithms	623	
20.11.1	McNemar's Test	623	
20.11.2	K-Fold Cross-Validated Paired t Test	623	
20.11.3	5 × 2 cv Paired t Test	624	
20.11.4	5 × 2 cv Paired F Test	625	

20.12	Comparing Multiple Algorithms: Analysis of Variance	626
20.13	Comparison over Multiple Datasets	630
20.13.1	Comparing Two Algorithms	631
20.13.2	Multiple Algorithms	633
20.14	Multivariate Tests	634
20.14.1	Comparing Two Algorithms	635
20.14.2	Comparing Multiple Algorithms	636
20.15	Notes	637
20.16	Exercises	638
20.17	References	640
A	Probability	643
A.1	Elements of Probability	643
A.1.1	Axioms of Probability	644
A.1.2	Conditional Probability	644
A.2	Random Variables	645
A.2.1	Probability Distribution and Density Functions	645
A.2.2	Joint Distribution and Density Functions	646
A.2.3	Conditional Distributions	646
A.2.4	Bayes' Rule	647
A.2.5	Expectation	647
A.2.6	Variance	648
A.2.7	Weak Law of Large Numbers	649
A.3	Special Random Variables	649
A.3.1	Bernoulli Distribution	649
A.3.2	Binomial Distribution	650
A.3.3	Multinomial Distribution	650
A.3.4	Uniform Distribution	650
A.3.5	Normal (Gaussian) Distribution	651
A.3.6	Chi-Square Distribution	652
A.3.7	t Distribution	653
A.3.8	F Distribution	653
A.4	References	653
B	Linear Algebra	655
B.1	Vectors	655
B.2	Matrices	657
B.3	Similarity of Vectors	658
B.4	Square Matrices	659

B.5	Linear Dependence and Ranks	659
B.6	Inverses	660
B.7	Positive Definite Matrices	660
B.8	Trace and Determinant	660
B.9	Eigenvalues and Eigenvectors	661
B.10	Spectral Decomposition	662
B.11	Singular Value Decomposition	662
B.12	References	663
C	Optimization	665
C.1	Introduction	665
C.2	Linear Optimization	667
C.3	Convex Optimization	667
C.4	Duality	668
C.5	Local Optimization	670
C.6	References	671
Index		673