

Contents

	<i>Preface</i>	page xii
	<i>Acknowledgements</i>	xxii
1	Introduction to Machine Learning	1
	1.1 Introduction	1
	1.2 Distinguishing Cats from Dogs: a Machine Learning Approach	1
	1.3 The Basic Taxonomy of Machine Learning Problems	6
	1.4 Mathematical Optimization	16
	1.5 Conclusion	18
	Part I Mathematical Optimization	19
2	Zero-Order Optimization Techniques	21
	2.1 Introduction	21
	2.2 The Zero-Order Optimality Condition	23
	2.3 Global Optimization Methods	24
	2.4 Local Optimization Methods	27
	2.5 Random Search	31
	2.6 Coordinate Search and Descent	39
	2.7 Conclusion	40
	2.8 Exercises	42
3	First-Order Optimization Techniques	45
	3.1 Introduction	45
	3.2 The First-Order Optimality Condition	45
	3.3 The Geometry of First-Order Taylor Series	52
	3.4 Computing Gradients Efficiently	55
	3.5 Gradient Descent	56
	3.6 Two Natural Weaknesses of Gradient Descent	65
	3.7 Conclusion	71
	3.8 Exercises	71
4	Second-Order Optimization Techniques	75
	4.1 The Second-Order Optimality Condition	75

4.2	The Geometry of Second-Order Taylor Series	78
4.3	Newton's Method	81
4.4	Two Natural Weaknesses of Newton's Method	90
4.5	Conclusion	91
4.6	Exercises	92
Part II	Linear Learning	97
5	Linear Regression	99
5.1	Introduction	99
5.2	Least Squares Linear Regression	99
5.3	Least Absolute Deviations	108
5.4	Regression Quality Metrics	111
5.5	Weighted Regression	113
5.6	Multi-Output Regression	116
5.7	Conclusion	120
5.8	Exercises	121
5.9	Endnotes	124
6	Linear Two-Class Classification	125
6.1	Introduction	125
6.2	Logistic Regression and the Cross Entropy Cost	125
6.3	Logistic Regression and the Softmax Cost	135
6.4	The Perceptron	140
6.5	Support Vector Machines	150
6.6	Which Approach Produces the Best Results?	157
6.7	The Categorical Cross Entropy Cost	158
6.8	Classification Quality Metrics	160
6.9	Weighted Two-Class Classification	167
6.10	Conclusion	170
6.11	Exercises	171
7	Linear Multi-Class Classification	174
7.1	Introduction	174
7.2	One-versus-All Multi-Class Classification	174
7.3	Multi-Class Classification and the Perceptron	184
7.4	Which Approach Produces the Best Results?	192
7.5	The Categorical Cross Entropy Cost Function	193
7.6	Classification Quality Metrics	198
7.7	Weighted Multi-Class Classification	202
7.8	Stochastic and Mini-Batch Learning	203
7.9	Conclusion	205
7.10	Exercises	205

8	Linear Unsupervised Learning	208
	8.1 Introduction	208
	8.2 Fixed Spanning Sets, Orthonormality, and Projections	208
	8.3 The Linear Autoencoder and Principal Component Analysis	213
	8.4 Recommender Systems	219
	8.5 K-Means Clustering	221
	8.6 General Matrix Factorization Techniques	227
	8.7 Conclusion	230
	8.8 Exercises	231
	8.9 Endnotes	233
9	Feature Engineering and Selection	237
	9.1 Introduction	237
	9.2 Histogram Features	238
	9.3 Feature Scaling via Standard Normalization	249
	9.4 Imputing Missing Values in a Dataset	254
	9.5 Feature Scaling via PCA-Sphering	255
	9.6 Feature Selection via Boosting	258
	9.7 Feature Selection via Regularization	264
	9.8 Conclusion	268
	9.9 Exercises	269
Part III	Nonlinear Learning	273
10	Principles of Nonlinear Feature Engineering	275
	10.1 Introduction	275
	10.2 Nonlinear Regression	275
	10.3 Nonlinear Multi-Output Regression	282
	10.4 Nonlinear Two-Class Classification	286
	10.5 Nonlinear Multi-Class Classification	290
	10.6 Nonlinear Unsupervised Learning	294
	10.7 Conclusion	298
	10.8 Exercises	298
11	Principles of Feature Learning	304
	11.1 Introduction	304
	11.2 Universal Approximators	307
	11.3 Universal Approximation of Real Data	323
	11.4 Naive Cross-Validation	335
	11.5 Efficient Cross-Validation via Boosting	340
	11.6 Efficient Cross-Validation via Regularization	350
	11.7 Testing Data	361
	11.8 Which Universal Approximator Works Best in Practice?	365
	11.9 Bagging Cross-Validated Models	366

x	Contents	
	11.10 K-Fold Cross-Validation	373
	11.11 When Feature Learning Fails	378
	11.12 Conclusion	379
	11.13 Exercises	380
12	Kernel Methods	383
	12.1 Introduction	383
	12.2 Fixed-Shape Universal Approximators	383
	12.3 The Kernel Trick	386
	12.4 Kernels as Measures of Similarity	396
	12.5 Optimization of Kernelized Models	397
	12.6 Cross-Validating Kernelized Learners	398
	12.7 Conclusion	399
	12.8 Exercises	399
13	Fully Connected Neural Networks	403
	13.1 Introduction	403
	13.2 Fully Connected Neural Networks	403
	13.3 Activation Functions	424
	13.4 The Backpropagation Algorithm	427
	13.5 Optimization of Neural Network Models	428
	13.6 Batch Normalization	430
	13.7 Cross-Validation via Early Stopping	438
	13.8 Conclusion	440
	13.9 Exercises	441
14	Tree-Based Learners	443
	14.1 Introduction	443
	14.2 From Stumps to Deep Trees	443
	14.3 Regression Trees	446
	14.4 Classification Trees	452
	14.5 Gradient Boosting	458
	14.6 Random Forests	462
	14.7 Cross-Validation Techniques for Recursively Defined Trees	464
	14.8 Conclusion	467
	14.9 Exercises	467
	Part IV Appendices	471
Appendix A	Advanced First- and Second-Order Optimization Methods	473
	A.1 Introduction	473
	A.2 Momentum-Accelerated Gradient Descent	473
	A.3 Normalized Gradient Descent	478
	A.4 Advanced Gradient-Based Methods	485

A.5	Mini-Batch Optimization	487
A.6	Conservative Steplength Rules	490
A.7	Newton’s Method, Regularization, and Nonconvex Functions	499
A.8	Hessian-Free Methods	502
Appendix B	Derivatives and Automatic Differentiation	511
B.1	Introduction	511
B.2	The Derivative	511
B.3	Derivative Rules for Elementary Functions and Operations	514
B.4	The Gradient	516
B.5	The Computation Graph	517
B.6	The Forward Mode of Automatic Differentiation	520
B.7	The Reverse Mode of Automatic Differentiation	526
B.8	Higher-Order Derivatives	529
B.9	Taylor Series	531
B.10	Using the autograd Library	536
Appendix C	Linear Algebra	546
C.1	Introduction	546
C.2	Vectors and Vector Operations	546
C.3	Matrices and Matrix Operations	553
C.4	Eigenvalues and Eigenvectors	556
C.5	Vector and Matrix Norms	559
	<i>References</i>	564
	<i>Index</i>	569